PhD degree in Molecular Medicine (curriculum in Computational Biology)

European School of Molecular Medicine (SEMM),

University of Milan and University of Naples "Federico II"

Settore disciplinare: Med/04

Detection of structural variations during liver cancer progression

Shruti Sinha

IEO, Milan

Matricola n. R09413

Supervisor: Dr. Francesca Ciccarelli

IEO, Milan

Added Supervisor: Dr. Gioacchino Natoli

IEO, Milan

Anno accademico 2013-2014

European School of Molecular Medicine

Thesis submitted by

SHRUTI SINHA

For the PhD in

MOLECULAR MEDICINE
(Curriculum in Computational Biology)

Thesis Title

Detection of structural variations during liver cancer progression

Supervising team

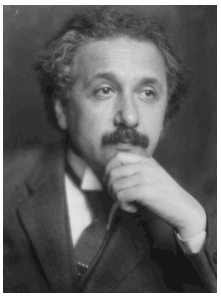
Supervisor	**Dr. Francesca Ciccarelli**
Internal advisor	**Dr. Gioacchino Natoli**
External advisor	**Dr. Tomás Marquès-Bonet**

Thesis approved by the supervisor

supervisor's signature

This thesis is dedicated to the world of science.

A small step towards exploring the endless possibilities.



"Albert Einstein portrait" by E. O. Hoppe. (1878-1972) Published on LIFE

To raise new questions, new possibilities, to regard old problems from a new angle, requires creative imagination and marks real advance in science.

- Albert Einstein

Table of contents

List of abbreviations

ABCB11: ATP-binding cassette, sub-family B member 11

ABCB4: ATP-binding cassette, sub-family B member 4

aCGH: Array comparative genomic hybridization

BSEP: Bile salt export pump

CBS: Circular binary segmentation

CIN: Chromosomal instability

CN-LOH: Copy neutral loss of heterozygosity

CNVs: Copy number variations

DGV: Database of Genomic Variants

DNA: Deoxyribonucleic acid

FFPE: Formalin-fixed paraffin embedded

FoSTeS: Fork stalling and template switching

HBV: Hepatitis B virus

HCC: Hepatocellular carcinoma

HCV: Hepatitis C virus

Indels: Insertion and deletions

JNKs: c-Jun NH2-terminal kinases

KO: Knock out

LCR: Low-copy repeat

LOH: Loss of heterozygosity

PCR: Polymerase chain reaction

MAPK: Mitogen-activated protein kinase

Mdr2: Multidrug resistance 2

MMBIR: Microhomology-mediated break-induced replication

MMEJ: Microhomology mediated end-joining

NAHR: Non-allelic homologous recombination

NGS: Next generation sequencing

NHEJ: Non-homologous end joining

PFIC: Progressive familial intrahepatic cholestasis

RNI: Reactive nitrogen intermediates

ROS: Reactive oxygen species

SNP: Single nucleotide polymorphism

SNVs: Single nucleotide variants

WES: Whole exome sequencing

WGS: Whole genome sequencing

List of figures

List of tables

Abstract

Hepatocellular carcinoma (HCC) is one of the most lethal cancers in the world and accounts for the vast majority of all liver cancers. HCC develops in response to various factors including viral infections, aflatoxin, alcohol and metabolic diseases. Recent studies have highlighted substantial differences in the acquired genomic alterations depending on the causative agent. Despite such a mutagen-dependent genetic heterogeneity, HCC is almost invariably associated with an underlying inflammatory state, whose direct contribution to the acquisition of critical genomic changes is not yet clear.

The aim of my PhD project has been to understand how chronic inflammation and fibrosis affect the cancer genome. We mapped the acquired genomic alterations in human and mouse HCCs induced by defects in hepatocyte biliary transporters. These HCCs arise as a result of chronic exposure to non-neutralized bile acids that cause the onset of chronic inflammation and develop into cancer in the absence of exogenous direct (viruses) or indirect (alcohol) mutagens. We first studied the mutational landscapes of human and mouse cancer genomes and found a surprisingly low number of somatic point mutations with no impairment of cancer genes. We next studied the acquisition of somatic copy number variations (CNVs) and used well-established approaches for detecting CNVs from SNP arrays and whole genome sequencing data. We also developed a novel method, GeneCNV, for the identification of CNVs from targeted re-sequencing screenings. Overall, we observed the acquisition of massive gene copy number gains and rearrangements in both human and mouse HCCs. Amplifications preferentially occurred at late stages of cancer development and frequently targeted the mitogen-activated protein kinase (MAPK) signalling pathway, in particular, direct regulators of c-Jun NH2-terminal kinases (JNKs). We showed that that pharmacological inhibition of JNK impairs the adenoma-to-carcinoma progression in mouse. This suggests that JNK inhibition may be a useful therapeutic approach to block HCC onset in bile salt export pump (BSEP) deficiency patients waiting for liver transplantation.

Altogether, this study showed that human BSEP-HCCs and mouse *Mdr2*-KO HCCs acquire a similar genomic signature, thus highlighting the remarkable analogy between human and mouse tumours with similar etiopathogenesis. This genomic signature differs from that of other HCCs profiled so far, which were for the most part virus induced. This demonstrates that HCC in the absence of external agents develops through genomic alterations that can be clearly distinguished from those determined by other etiological factors.

Introduction

1.1 Overview of human liver cancer

Liver cancer is the second leading cause of cancer-related deaths and the sixth most common cancer type in the world (Ferlay J 2012) (Figure 1). Primary liver cancers include hepatocellular carcinoma (HCC), intrahepatic bile duct carcinoma (cholangiocarcinoma), hepatoblastoma, bile duct cystadenocarcinoma, hemangiosarcoma and epithelioid hemangioendothelioma (Farazi and DePinho 2006). Among primary liver cancers, HCC is the most common tumour and accounts for 70-85% of all liver cancers (Perz, et al. 2006). HCC is usually associated with high mortality due to unresponsiveness to treatment, and tumour relapse after partial hepatectomy (Block, et al. 2003; Llovet 2005).

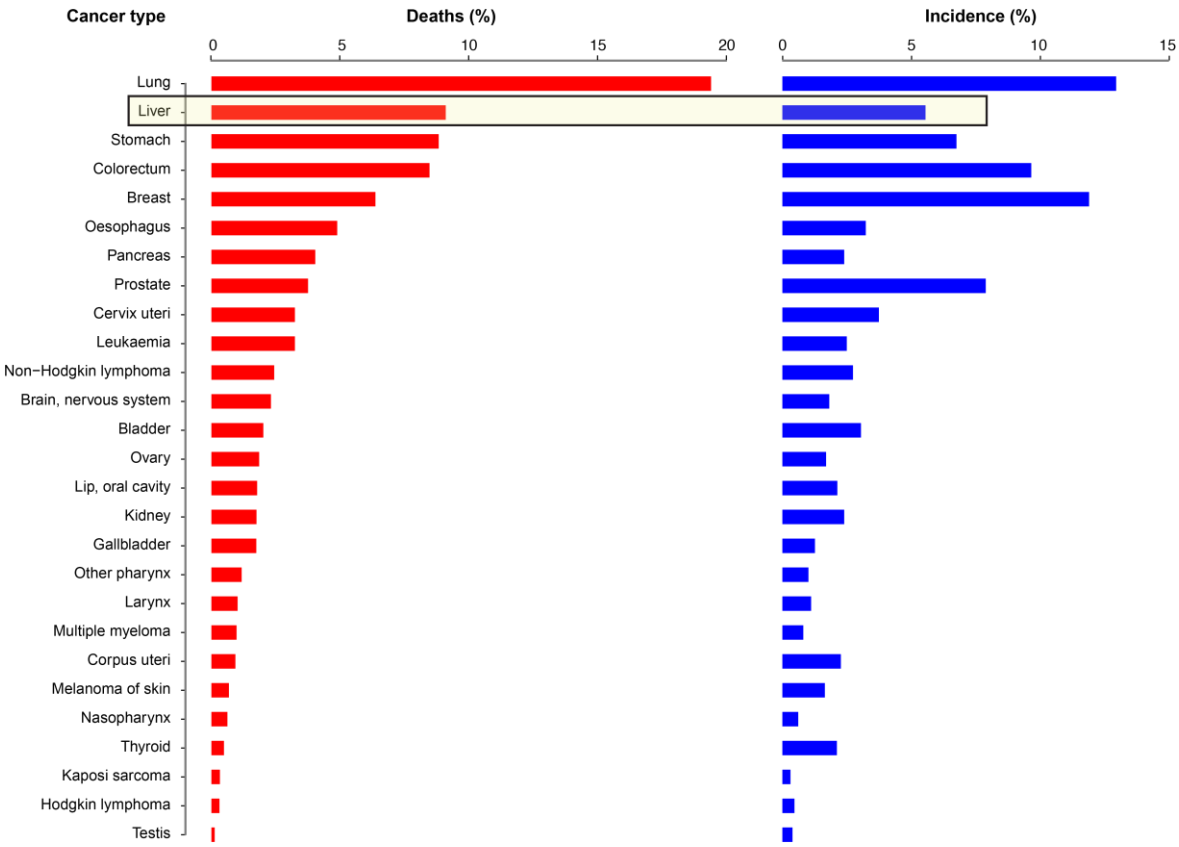


Figure 1: Percentage of deaths and incidence by cancers worldwide

For each cancer type, the percentage of deaths over the total deaths by cancer and the percentage of incidence over the total incidence of cancers worldwide are reported. Original data is taken from GLOBOCAN 2012 (Ferlay J 2012), which excludes non-melanoma skin cancer from its study

because of the difficulties in collecting (and counting) such tumours (http://globocan.iarc.fr/Pages/cancer.aspx).

HCC may arise in response to exposure to external factors such as virus infection, aflatoxin and alcohol, or by metabolic diseases including obesity and diabetes (Block, et al. 2003; El-Serag and Rudolph 2007). The most common cause of HCC in adults is chronic viral infections due to hepatitis B (HBV) and hepatitis C (HCV) that account for 78% of HCCs (Block, et al. 2003) (Figure 2).

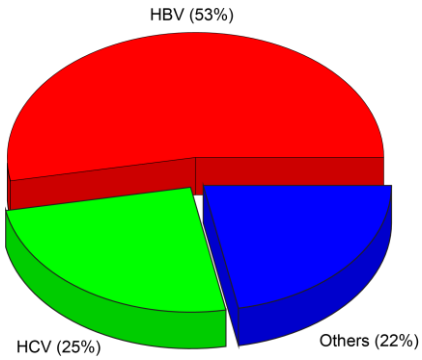


Figure 2: Distribution of aetiologies in HCCs

Shown are the percentages of HCCs associated with the virus infections as compared to other causes. Taken from (Block, et al. 2003).

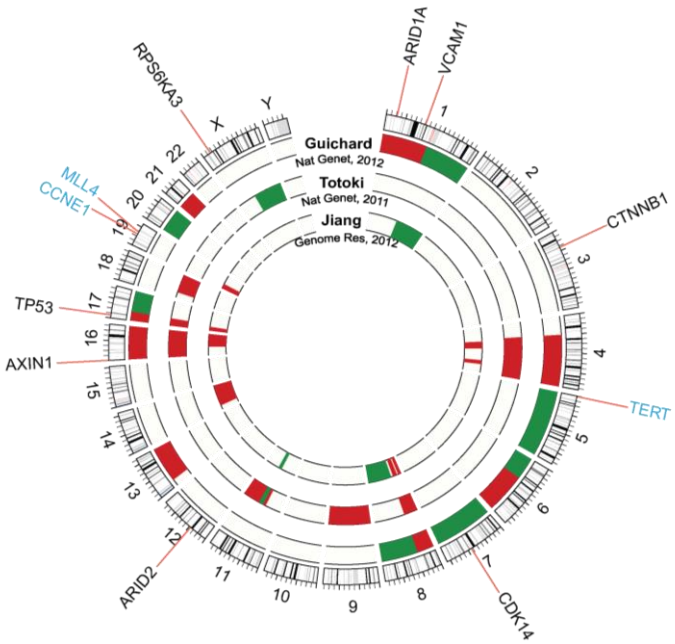
1.2 Genomic landscape of hepatocellular carcinoma

Genomic re-sequencing of human HCCs has highlighted recurrent mutations in key cancer genes such as *TP53*, *CTNNB1*, and chromatin regulators (Li, Zhao, et al. 2011; Totoki, et al. 2011; Fujimoto, et al. 2012; Guichard, et al. 2012; Huang, et al. 2012; Jiang, et al. 2012; Sung, et al. 2012) (Table 1, Figure 3). These studies have also reported genomic regions that undergo frequent copy number variations (CNVs) such as amplifications of 11q1, 8q and deletions of 4q, 8p and 17p (Totoki, et al. 2011; Guichard, et al. 2012; Jiang, et al. 2012) (Figure 3).

Table 1: Genomic re-sequencing studies of HCCs

Screening	*Screening Type*	*Patients*	*Mutated genes*	*Non Synonymous mutations*	*Mutation frequency (mutation /Mbp)*
Guichard, Nature Genetics 2012	WES	24	872	1,056	1.02
Huang, Nature Genetics 2012	WES	10	347	356	0.99
Li, Nature Genetics 2011	WES	10	411	429	0.98
Fujimoto, Nature Genetics 2012	WGS	25	206	1,996	4.2
Jiang, Genome Research 2012	WGS	4	221	228	3.8
Totoki, Nature Genetics 2011	WGS	1	71	72	4.11
Sung, Nature Genetics 2012	WGS	88	NA	NA	NA

For each screening reported are the type of screening performed (WES=whole exome sequencing, WGS=whole genome sequencing), number of patients analysed, number of mutated genes, number of nonsynonymous mutations and average mutation frequency in each screening. Mutation frequency in whole exome sequencing is calculated as the number of nonsynonymous mutation over the total bases targeted for sequencing. In case of whole genome sequencing the mutation frequency is computed as the number of mutations over the total size of the genome.



Genes	*Screenings*
ARID1A	Huang Nat Genet (2012)
	Fujimoto Nat Genet (2012)
	Guichard Nat Genet (2012)
VCAM1	Huang Nat Genet (2012)
CTNNB1	Huang Nat Genet (2012)
	Fujimoto Nat Genet (2012)
	Guichard Nat Genet (2012)
	Jiang Genome Res (2012)
	Totoki Nat Genet (2011)
	Li Nat Genet (2011)
TERT	Fujimoto Nat Genet (2012)
	Sung Nat. Genet (2012)
CDK14	Huang Nat Genet (2012)
ARID2	Fujimoto Nat Genet (2012)
	Guichard Nat Genet (2012)
AXIN1	Guichard Nat Genet (2012)
TP53	Huang Nat Genet (2012)
	Fujimoto Nat Genet (2012)
	Guichard Nat Genet (2012)
	Jiang Genome Res (2012)
	Totoki Nat Genet (2011)
	Li Nat Genet (2011)
	Sung Nat. Genet (2012)
CCNE1	Sung Nat. Genet (2012)
MLL4	Sung Nat. Genet (2012)
RPS6KA3	Guichard Nat Genet (2012)

Figure 3: Frequently mutated genes and CNV regions in HCCs

Shown are the frequently mutated genes (Zhang 2012) and CNV regions reported in the HCC studies (Table 1). Frequent HBV integration target genes are shown in blue.

These studies highlighted substantial differences in the acquired genomic alterations across HCC samples, thus suggesting genetic heterogeneity of HCC depending on the initiating agents. For example, genes encoding components of the chromatin-remodelling complexes are frequently mutated in hepatitis HCV- but not in HBV-associated HCC (Li, Zhao, et al. 2011). They also revealed a strong dependence of the acquired mutation signature on the underlying mutagenic mechanism (Zhang 2012). For example, the exposure to different genotoxic chemicals produces distinct mutation patterns, namely C:G to A:T transversions in the case of aflatoxin (Fujimoto, et al. 2012; Guichard, et al. 2012; Huang, et al. 2012), and T:A to A:T transversions in the case of aristolochic acid and vinyl chloride (Huang, et al. 2012). Moreover, the integration of the viral DNA in the host genome has been observed in HBV- but not in HCV-associated HCCs (Brechot, et al. 1980; Fujimoto, et al. 2012; Jiang, et al. 2012; Sung, et al. 2012). These data clearly show that HCC has a complex pathogenesis, which translates into the heterogeneous genomic landscape of these tumours as a consequence to mutagens, inflammation and sustained regeneration that all cooperate to promote cancer.

1.3 External factors as oncogenic events of hepatocellular carcinoma

HCC is a multistep process that involves the progressive accumulation of different genetic alterations as a response to mutagens including hepatitis B and C viral infections, aflatoxin, alcohol or metabolic diseases. The viral-associated mechanisms driving liver cancer are complex. Hepatitis B and C viruses (HBV and HCV, respectively) may help in development of cancer either through direct involvement in the transformation or in

indirect ways by triggering immune response of the host. HBV infection has been reported to promote carcinogenesis through three mechanisms: (i) genomic instability or alteration in the function of genes by integration of HBV DNA into the host genome; (ii) modulation of cell proliferation and viability due to expression of viral proteins; and (iii) increase in genetic damage due to hepatic inflammation caused by virus specific T cells (Sung, et al. 2012). The direct involvement of HBV in HCC development has been supported by its DNA integration in the host genome. DNA integration of HBV in host genome has been associated with microdeletions affecting cancer-related genes like telomerase reverse transcriptase (*TERT*), platelet-derived-growth-factor receptor (*PDGFRβ*), platelet-derived-growth-factor (*PDGFβ*), mitogen activated protein kinase 1 (*MAPK1*), and the activation of proto-oncogenes of the myc family, predominantly the *N-myc2* oncogene (Farazi and DePinho 2006; Kremsdorf, et al. 2006). Additionally, transcriptional activity of hepatitis protein x (HBx) has been reported to bind and inactivate p53, resulting in increased cellular proliferation and survival (Ueda, et al. 1995). HBx also transactivates many growth control genes such as SRC tyrosine kinases, Ras, Raf, MAPK, ERK, JNK and others (Farazi and DePinho 2006).

Unlike HBV, HCV does not integrate in the host DNA. HCV alters host genes expression and cellular phenotypes by expression of viral proteins (Levrero 2006) or due to inflammatory responses to the oxidative stress as a result of persistent infection (McGivern and Lemon 2011). The core proteins and non-structural (NS) proteins, NS3 and NS5A have been reported to directly contribute to the oncogenic transformation (Arzumanyan, et al. 2013). HCV core proteins have been shown to interact with components of the MAPK signalling pathway (such as ERK, MEK and Raf) and NS5A has been shown to inactivate p53 by sequestration to the perinuclear membrane (Farazi and DePinho 2006).

Aflatoxin B1 is a mycotoxin produced by fungi: *Aspergillus flavus* and *Aspergillus parasiticus*. It usually causes base mutations that may lead to DNA break and oxidative damage (Hussain, et al. 2007; Hamid, et al. 2013). For example, it has been shown to have

positive association with G:C to T:A transversion in the codon 249 of *TP53* (249^{ser}, arginine to serine) (Hussain, et al. 2007). Other external factors such as alcohol and metabolic diseases are often associated with oxidative stress and cirrhosis (Seitz and Stickel 2007; Marra, et al. 2008).

Regardless of the initiating agent, HCC is believed to occur due to increased liver cell turnover, induced by chronic liver injury and regeneration, in a context of inflammation and oxidative DNA damage (Levrero 2006) (Figure 4).

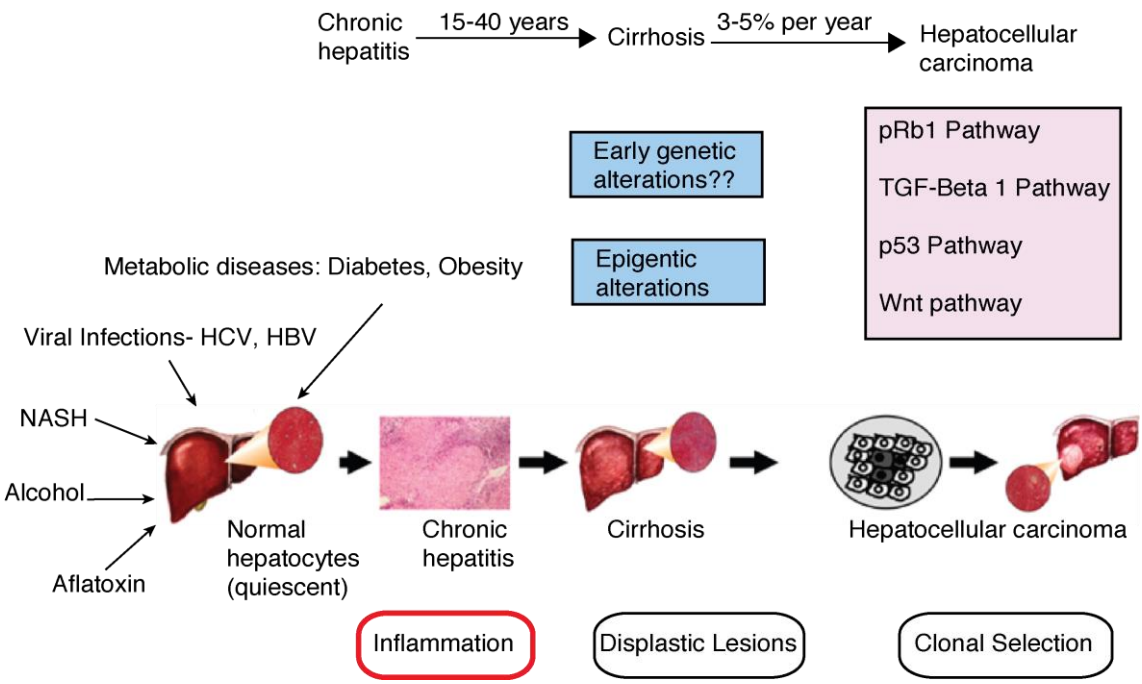


Figure 4: Pathogenesis of HCC

Shown are the etiological agents and known pathways regulating cell proliferation or apoptosis that are affected in HCCs. Dysplastic nodules and macroregenerative nodules are considered as pre-neoplastic lesions. All HCCs, irrespective of the etiological agent are secondary to inflammation (highlighted in red). Taken from (Levrero 2006).

1.4 Inflammation and cancers

Chronic inflammation has been causally associated with cancer by Rudolf Virchow as early as in the 19th century, and is nowadays acknowledged as one of the major risk

factors for malignancies (Balkwill and Mantovani 2001). The inflammatory state of tumour cells is one of the enabling characteristics that confer the acquisition of alterations in the eight hallmarks of cancer (Hanahan and Weinberg 2000, 2011) (Figure 5). Chronic inflammation is linked to a variety of cancers including liver, colon, gastric, bladder, cervical, oesophageal, ovarian, prostate and thyroid cancers (Mantovani, et al. 2008). Inflammatory components (cells, cytokines and chemokines) infiltrate the microenvironment of most tumour lesions and have tumour-promoting effects (Mantovani, et al. 2008). As a consequence, several anti-inflammatory drugs have been reported to prevent tumour onset or to delay tumour progression and are currently used to treat cancer (Ulrich, et al. 2006). Furthermore, many interesting insights have been uncovered on the molecular events linking chronic inflammation to tumour formation and development. The emerging picture suggests a feedback loop between tumours and inflammatory responses. Activated innate immune cells stimulate tumour growth and progression via secretion of cytokines and pro-inflammatory mediators. Cancer cells in return produce soluble mediators that again recruit and activate inflammatory cells, further promoting tumour and thus creating a feedback loop (Grivennikov, et al. 2010).

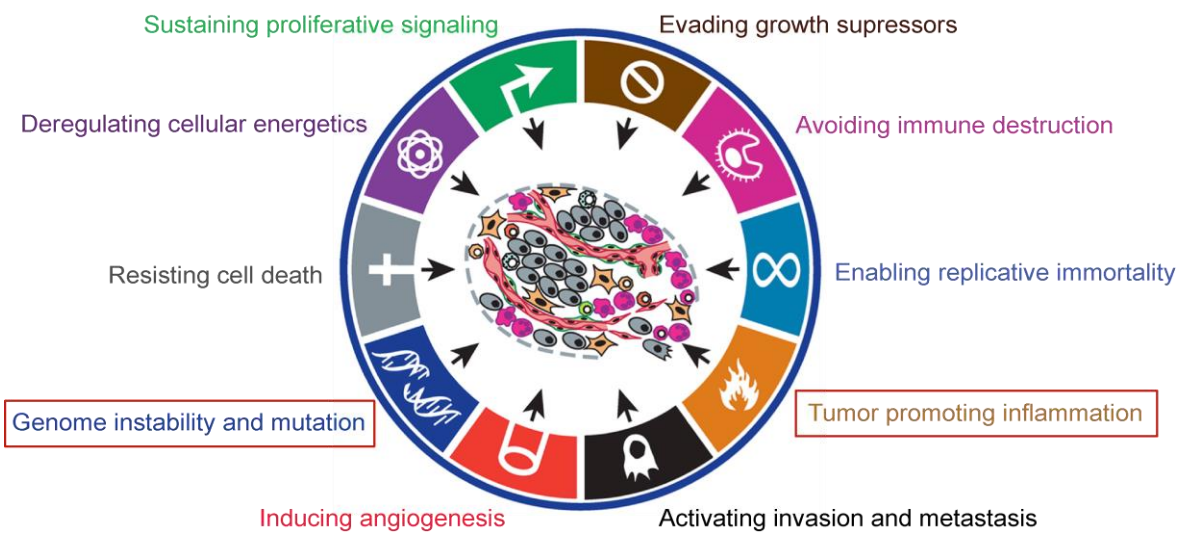


Figure 5: Hallmarks of cancer

Tumour cells are characterized by the acquisition of eight functional capabilities: self-sufficiency in growth signals, insensitivity to antigrowth signals, evasion from apoptosis, limitless replicative potential, sustained angiogenesis, tissue invasion and metastasis, reprogramming the energy metabolism and evading immune suppression. Genome instability and chronic inflammation are enabling characteristics that promote the tumour progression. Taken from (Hanahan and Weinberg 2000, 2011).

Despite its primary role, a clear evidence of a direct mutagenic potential of inflammation in cancer is still missing. An inflammatory microenvironment has been proposed to increase DNA damage leading to genomic instability through reactive oxygen species (ROS) and reactive nitrogen intermediates (RNI) (Elinav, et al. 2013). ROS and RNI are activated by macrophages and neutrophils or are induced intracellularly in pre-malignant cells by inflammatory cytokines (Mantovani, et al. 2008). They then react with DNA packed into chromatin after having diffused through the extracellular matrix, penetrated a cell, crossed its cytoplasm, and entered the nucleus. The stability and longevity of ROS and RNI released by inflammatory cells and their ability to react with chromatin is still a matter of controversy (Grivennikov, et al. 2010).

1.4.1 Role of inflammation in hepatocellular carcinoma

The majority of HCCs arise in the background of chronic liver injury including chronic hepatitis and cirrhosis (Laurent-Puig and Zucman-Rossi 2006). Recent studies have indicated that the inflammatory reaction is characteristic of chronic liver injury and actively participates in the development of hepatic fibrosis, as well as in the activation of the potent regenerative response of liver parenchyma. Inflammation-related signalling pathways have been implicated in liver tumour initiation and progression. These include

transcription factors of the NF-κB family, signal transducer and activator of transcription 3 (STAT3), cytokines such as TNF-α, IL-6 and IL-1a and ligands of the epidermal growth factor receptor (EGFR) (Berasain, et al. 2009; Alison, et al. 2011) (Figure 6).

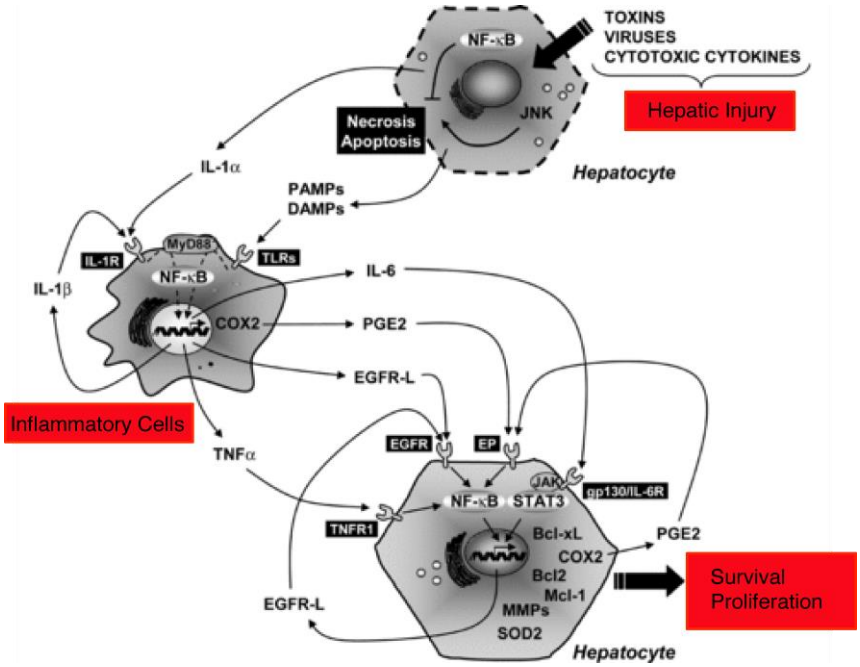


Figure 6: Molecular pathways linking inflammation and liver cancer

Shown are the molecules and signalling pathways that have been associated with liver cirrhosis and HCC. The critical components linking inflammation and liver cancer are the transcription factors of the NF-κB family, signal transducer and activator of transcription 3 (STAT3), cytokines such as TNF-α, IL-6 and IL-1a and ligands of the epidermal growth factor receptor (EGFR). Taken from (Berasain, et al. 2009).

TNF-α plays an important role in the promotion of HCC in mice induced by choline-deficient and ethionine-supplemented diet and also in *Mdr2*-KO mice, a genetic mouse model for inflammation related HCC (Berasain, et al. 2009). NF-κB is known to play dual role in the development of liver cancer. Its activation has been associated in the liver during neoplastic transformation whereas its hepatocyte-specific inactivation has been shown to lead to higher incidence of liver tumours upon treatment with the carcinogen DEN (Maeda, et al. 2005). In addition, loss of hepatocyte NF-κB activity enhances

chemical carcinogenesis through sustained c-Jun-N-terminal kinase 1 (JNK1) activation (Berasain, et al. 2009). High circulating IL-6 levels have been observed in inflammatory conditions of the liver, including chronic alcohol consumption, viral infections, or hepatic iron accumulation in HCC patients (Naugler and Karin 2008). Furthermore, IL-6 binding triggers JAK-STAT3 signalling pathway, which is enhanced in hepatic inflammation and HCC (Berasain, et al. 2009). Similar to cytokines, activation of growth factors expression and signalling such as insulin-like growth factor (IGF), hepatocyte growth factor (HGF), Wingless (Wnt), transforming growth factor beta (TGFβ), and the EGFR signalling system is frequently observed in chronic inflammatory liver diseases and HCC (Breuhahn, et al. 2006). Though signalling pathways related to inflammation have been liked to tumour initiation and progression in liver cancer, the direct potential of inflammation in inducing genomic alterations is still unknown.

1.5 Hepatocellular carcinoma induced by defects in biliary transporter genes

Bile is a vital secretion of liver and is essential for intestinal digestion, absorption of lipids and elimination of environmental toxins, carcinogens, drugs, and their metabolites. Formation and flux of bile in the liver are maintained by distinct transport systems: basolateral (sinusoidal) transport and apical (canalicular) transport (Pauli-Magnus and Meier 2005) (Figure 7).

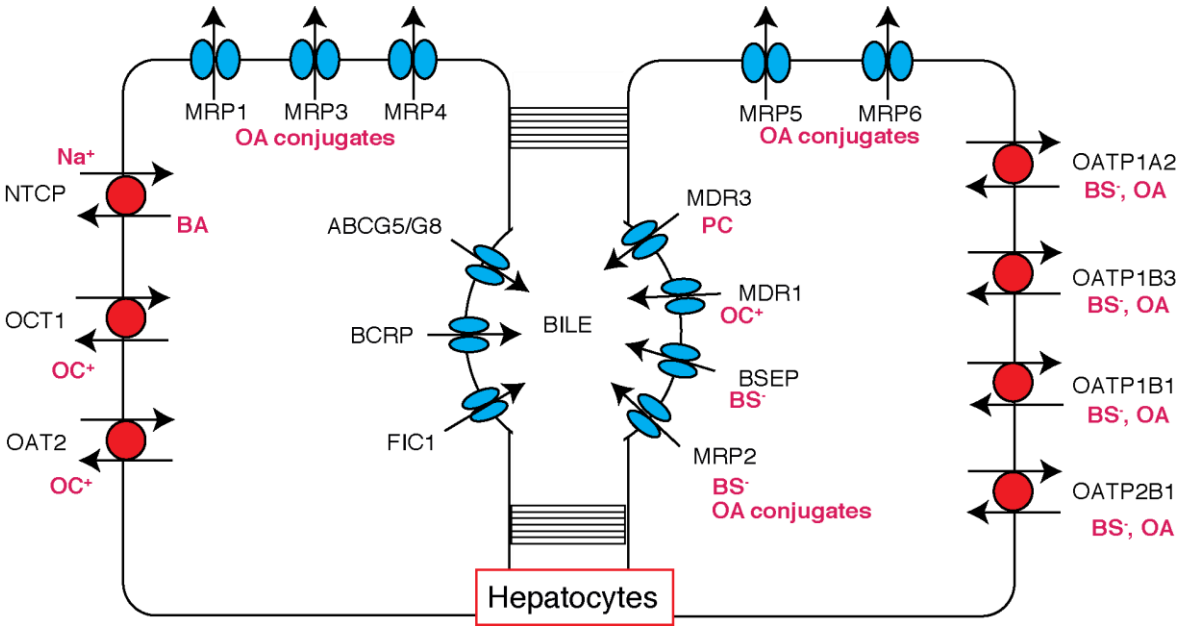


Figure 7: Hepatocyte bile transporters

Shown are the hepatobiliary transporter systems. Basolateral transport system includes MRP1, MRP3, MRP4, MRP5, MRP6, NTCP, OATP1A2, OATP1B3, OATP1B1, OATP2B1, OCT1 and OAT2. Apical transport system includes MRP2, ABCG5, ABCG8, BSEP, MDR1, MDR3, BCRP and FIC1. Efflux transporters are represented in blue and uptake transporters are represented in red. Bile salts (BS^-) are taken via basolateral transporters (NTCP) and organic anions by OATPs (OATP1A2, OATP1B3, OATP1B1 and OATP2B1). MDR3 translocates phosphatidylcholine from the inner to the outer leaflet. BSEP excrete the monovalent bile salts while MRP2 excretes organic anion conjugates and divalent bile salts. Other MRPs: MRP1, MRP2, MRP4, MRP5 and MRP6 are involved in the efflux of organic anions. BS=bile salts; OC= organic conjugates; BA=bile acids; OA= organic anion. Taken from (Pauli-Magnus and Meier 2006).

Basolateral transporters are involved in hepatic uptake of endogenous and exogenous substances from sinusoidal blood plasma. So far, two key players have been identified: the sodium dependent pathway represented by the sodium-taurocholate cotransporting polypeptide NTCP (*SLC10A1*) and the sodium independent pathways represented by different members of the superfamily of organic anion transporting

polypeptides (OATP/SLCO) (Pauli-Magnus and Meier 2005) (Figure 7). NTCP uptake is mostly limited to conjugated bile salts and certain sulphated steroids. NTCP accounts for more than 80% of conjugated bile salts uptake in the liver (Pauli-Magnus and Meier 2006). The OATP superfamily comprises of 9 human members, many of which are present at different expression levels in liver (Hagenbuch and Meier 2004). These transporters are involved in the trafficking of several metabolites, conjugated and unconjugated bile salts, bromosulfophthalein, neutral steroids, steroid sulfates and glucuronides, and selected organic cations (Pauli-Magnus and Meier 2006). They are also involved in uptake of many drugs including antihistamine fexofenadine, opioid peptides, digoxin, the HMG CoA-reductase inhibitor pravastatin, the angiotensin-converting enzyme inhibitor enalapril, and the antimetabolite methotrexate. Finally, they aid in the uptake of hepatotoxins phalloidin, microcystin and amanitin. Basolateral system also localizes adenosine triphosphate (ATP)-dependent efflux pumps belonging to the multidrug resistance-associated proteins (MRPs) (*ABCC*): MRP1 (*ABCC1*), MRP3 (*ABCC3*), MRP4 (*ABCC4*), MRP5 (*ABCC5*), and MRP6 (*ABCC6*) (Pauli-Magnus and Meier 2005) (Figure 7). These proteins are multispecific transporters for different organic anions. All basolateral transporter genes are extensively regulated by transcriptional and posttranscriptional processes that allow adaptation to changes in response to the intracellular accumulation of bile salts (Pauli-Magnus and Meier 2006).

Apical (canalicular) transporters mediate the secretion of bile salts and other bile constituents across the canalicular membrane of hepatocytes. These are ATP-binding cassette transporters that use ATP to pump solutes into bile against steep concentration gradients. The members of ABC transporters that are expressed in liver include MRP2 (*ABCC2*), ABCG2 (*ABCG2*), ABCG5 (*ABCG5*) ABCG8 (*ABCG8*) MDR1 (*ABCB1*), MDR3 (*ABCB4*) and BSEP (*ABCB11*), (Table 2, Figure 7). Multi-drug resistance protein 2 (MRP2) is a member of multidrug resistance-associated proteins family that recognizes a wide spectrum of organic anions, including bilirubin-diglucuronide, glutathione

conjugates, leukotriene C4, and divalent bile salt conjugates. *ABCC2* encodes MRP2 and impairment of this gene due to inactivating mutations causes Dublin-Johnson syndrome (Kartenbeck, et al. 1996; Keitel, et al. 2000). ATP-binding cassette sub-family G member 5 (ABCG5) and ATP-binding cassette sub-family G member 8 (ABCG8) genes are involved in excretion of sterols and cholesterol (Figure 7). Mutations in these genes result in decreased biliary excretion causing sitosterolemia leading to hypercholesterolemia and premature atherosclerosis (Lu, et al. 2001) (Table 2). *ABCB1* encodes MDR1 and is involved in transporting amphipathic basic or cationic compounds (Figure 7). Though no known disease has been associated with defects in this gene in humans, *Abcb1* knockout mice are reported to be hypersensitive to many drugs and toxins (Schinkel, et al. 1995). P-glycoprotein 3 (MDR3) and the bile salt export pump (BSEP) are highly conserved members of multidrug resistance protein family (Pauli-Magnus and Meier 2005). They have been identified as the key players in the secretion of bile salts and other bile constituents across the canalicular membrane of hepatocytes (Pauli-Magnus, et al. 2004). *ABCB11* encodes BSEP, and mediates the cellular excretion of numerous conjugated bile salts such as taurine- or glycine-conjugated cholate, chenodeoxycholate, and deoxycholate and is the predominant bile salt efflux system of hepatocytes (Figure 7). MDR3 is encoded by *ABCB4*, which translocates phosphatidylcholine from the inner to the outer leaflet of the canalicular membrane (Figure 7). Another transporter, FIC (*ATP8B1*) has been reported to regulate the enterohepatic bile salt pool and eliminate the hydrophobic substances from the enterohepatic circulation (Pauli-Magnus and Meier 2005). Defects in the three genes (*ABCB4*, *ATP8B1* and *ABCB11*), involved in the regulation and formation of bile salts cause progressive familial intrahepatic cholestasis (PFIC) (Davit-Spraul, et al. 2009; Jacquemin 2012).

Table 2: Human canalicular transporter proteins and associated diseases

Name	*Gene*	*Chromosome*	*Transport Function*	*Disease*
MRP2	*ABCC2*	10q24	OA	Dublin-Johnson syndrome
ABCG2	*ABCG2*	4q22	Multispecific	-
ABCG5	*ABCG5*	2p21	Sterols	Sitosterolemia
ABCG8	*ABCG8*	2p21	Sterols	Sitosterolemia
MDR1	*ABCB1*	7q21	Multispecific	-
MDR3	*ABCB4*	7q21	PC	PFIC3
BSEP	*ABCB11*	2q24	BS	PFIC2

Reported are the names of the canalicular transporter proteins, gene name, location in the human genome, substrate for transporting function and the associated disease. Overexpression of MDR1 and ABCG2 confers resistance to drugs and inherited defects in the other transporter genes lead to diseased state. PC= phosphatidylcholine, BS= bile salts, OA=organic anions, PFIC=progressive familial intrahepatic cholestasis.

1.5.1 Bile salt export pump deficiency and paediatric hepatocellular carcinoma

PFIC designates a heterogeneous group of rare autosomal recessive disorders that usually appear in infancy or early childhood and manifest with hepatocellular damage and cholestasis due to defects in bile formation (Jacquemin 2012). Three types of PFIC (PFIC1, PFIC3 and PFIC2) have been identified and have been associated with inherited inactivating mutations in the hepatocyte membrane transporter genes *ATP8B1*, *ABCB4*, and *ABCB11* respectively (Davit-Spraul, et al. 2009; Jacquemin 2012) (Figure 8).

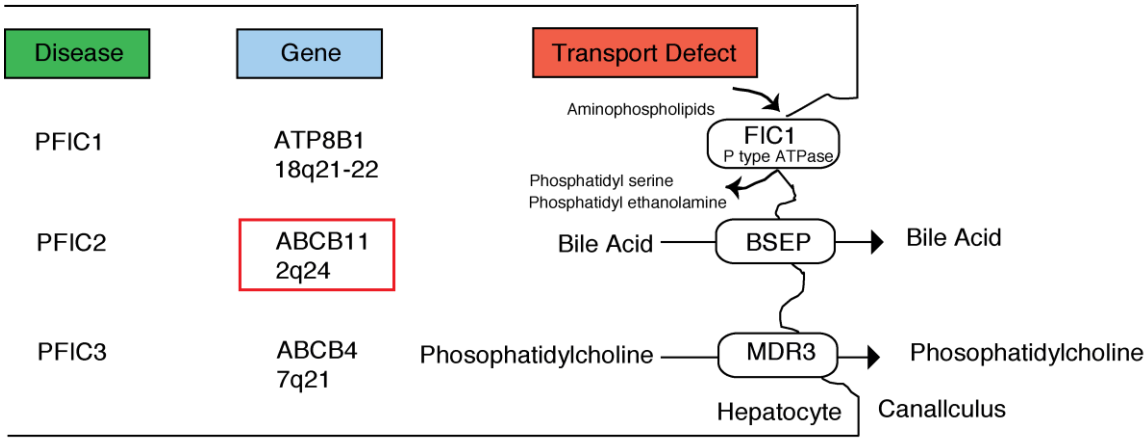


Figure 8: Progressive familial intrahepatic cholestasis

Shown are the types of PFIC and the corresponding mutated genes involved in its development along with the function of the gene in the liver. Impairment of *ABCB11* (highlighted in red) in humans and *ABCB4* ortholog in mouse leads to tumour development in the respective species. Taken from (Davit-Spraul, et al. 2009)

PFIC1 or "Byler disease" is caused by mutations in the *ATP8B1* gene (Figure 8). This gene encodes a P-type ATPase (FIC1) and is located on the canalicular membrane of hepatocytes (Figure 7). It is mainly expressed in cholangiocytes within the liver. The exact mechanism of how mutations in this gene cause cholestasis is unclear. It is assumed that the abnormal protein function of FIC1 could indirectly disrupt the biliary secretion of bile acids. This is in agreement with the low biliary bile acid concentrations and the chronic diarrhoea present in PFIC1 patients (Jacquemin 2012). Studies have also reported substantial down-regulation of the farnesoid X receptor (FXR) in PFIC1 patients with mutations in *ATP8B1*. FXR is a nuclear receptor, which is involved in the regulation of bile acid metabolism. Low expression level of FXR subsequently down-regulates BSEP in the liver and up-regulates bile acid synthesis and apical sodium bile salt transporter (ASBT) in the intestine (Alvarez, et al. 2004; Chen, et al. 2004) leading to hepatocyte bile acid overload (Jacquemin 2012).

PFIC2 is caused by inherited mutations in the *ABCB11* gene (Table 2, Figure 8). *ABCB11* gene encodes ATP-dependent canalicular BSEP, which is the major exporter of primary bile acids (Figure 7 and Figure 8). It is expressed at the hepatocyte canalicular membrane in human liver (Davit-Spraul, et al. 2009; Jacquemin 2012). Impairment of BSEP causes decrease in biliary bile salts secretion leading to decreased bile flow and accumulation of bile salts within the hepatocyte resulting in severe hepatocellular damage (Davit-Spraul, et al. 2010). BSEP deficiency represents a phenotypic continuum between PFIC2 and a milder form of the disease, known as benign recurrent intrahepatic cholestasis (BRIC2) (van Mil, et al. 2004).

PFIC3 represents an important example of canalicular transport defect leading to the development of cholangiopathy. It is caused by defects in *ABCB4* gene (Table 2, Figure 8). *ABCB4* encodes class III multidrug resistance (MDR3) of P-glycoprotein (P-gp) protein. MDR3 is a phospholipid translocator, which is involved in the excretion of biliary phospholipid and is predominantly expressed in the canalicular membrane of hepatocytes (Davit-Spraul, et al. 2010) (Figure 7 and Figure 8). PFIC3 is caused by injuries to bile canaliculi and biliary epithelium as a result of the toxicity of bile in which detergent bile salts are not inactivated by phospholipids. This suggests that the mechanism of liver damage in PFIC3 is most likely related to the absence of biliary phospholipids (Jacquemin 2012).

Patients with PFIC2 and PFIC3 may develop liver tumour and monitoring of liver cancer in these patients is usually advised (Jacquemin 2012). In particular, PFIC2 patients are at high risk of hepatobiliary malignancy (hepatocellular carcinoma or cholangiocarcinoma) (Jacquemin 2012). Fatal peripheral cholangiocarcinoma due to inherited mutations in *ABCB11* in the absence of BSEP expression was observed in some PFIC2 patients (Scheimann, et al. 2007). Another study on PFIC2 with severe BSEP deficiency reported presence of hepatocellular carcinoma or cholangiocarcinoma in 15% of the patients (Strautnieks, et al. 2008). In particular, two protein-truncating mutations in

BSEP conferred particular high risk (38%) to develop malignancy compared to less severe genotypes (Strautnieks, et al. 2008). These findings further support that BSEP deficiency is associated with increased susceptibility to hepatobiliary malignancy. BSEP deficiency is hypothesized to cause cholangiocarcinoma through bile-composition shifts or bile-acid damage within cells capable of hepatocytic/cholangiocytic differentiation (Scheimann, et al. 2007). Bile acids induce mutation via production of ROS and RNS (Bernstein, et al. 2005). Damage induced by intrahepatocytic accumulation of bile acids is hypothesized to lead to hepatocellular carcinoma (Knisely, et al. 2006). This suggests, that the loss of functional BSEP results in intrahepatocytic accumulation of bile acids causing liver injury resulting in chronic inflammation and to the early onset of hepatocellular carcinoma (Knisely, et al. 2006). Since HCCs caused by inactivating mutations in *ABCB11* occur in the background of chronic inflammation and fibrosis in absence of external mutagens, this liver cancer type provides the opportunity to understand the contribution of chronic inflammation and fibrosis to the acquired genomic modifications that trigger liver cancer in the absence of other mutagenic factors.

1.5.2 *Mdr2*-KO mouse model of hepatocellular carcinoma

Mdr2 is a bile transporter gene in mouse, which is an ortholog of human *ABCB4* gene. Homozygous disruption of *Mdr2* in mouse has been shown to result in nonsupportive inflammatory cholangitis and HCCs (Mauad, et al. 1994). *Mdr2*-KO mice lack the *Mdr2* encoded P-glycoprotein located on the canalicular membrane of hepatocytes and are therefore unable to secrete phosphatidylcholine into bile ducts. The resulting precipitation of bile salts induces hepatocellular damage, inflammation and eventually HCC with high penetrance (Smit, et al. 1993; Mauad, et al. 1994). In *Mdr2*-KO mice, HCC progresses through well-defined temporal stages starting with inflammation, which then develops into dysplasia (from 4 months of age), adenoma-like dysplastic nodules (7 months) and

eventually HCC starting from 10 months that is detectable in virtually 100% of mice by 16 months of age (Pikarsky, et al. 2004; Katzenellenbogen, et al. 2007) (Figure 9). Moreover, since HCCs arise within adenomas (Jang, et al. 1992), each neoplastic nodule offers a screenshot of the adenoma-to-carcinoma transition and displays a variable HCC fraction, which tends to increase with time.

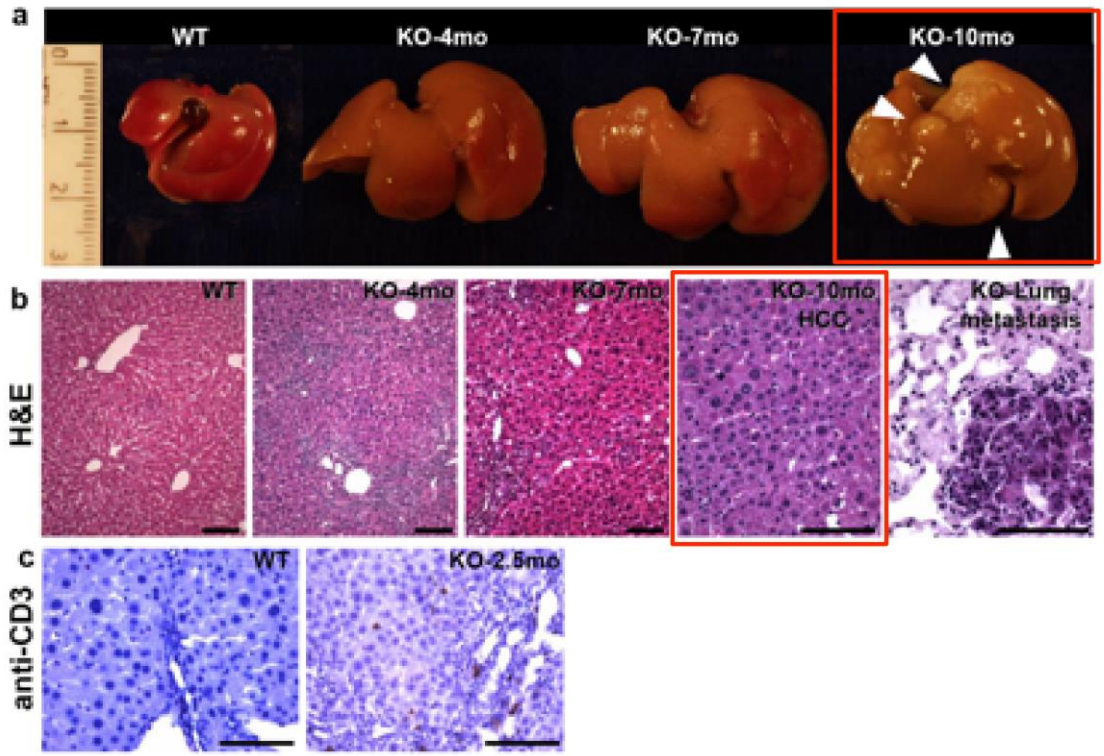


Figure 9: *Mdr2*-KO mice develop HCC on the background of chronic hepatitis

Shown are the livers of WT and KO mice sacrificed at the indicated time points. Arrowheads indicate tumours. Scale in cm (a). Hematoxylin and eosin (H&E) stained sections from livers from WT and KO mice of the indicated time points are shown. At 4 months, inflammation and ductular proliferation are prominent. At 7 months, there is severe architectural and cytologic dysplasia and HCC develops between 9-12 months (highlighted in red), which later disseminates metastases (b). The presence of T-cells in the mixed inflammatory infiltrate is highlighted by CD3 immunostaining (c). Scale Bars=50 mm. Taken from (Pikarsky, et al. 2004)

1.6 Genomic copy number alterations

1.6.1 Definition and mechanisms of formation

Copy number variations (CNVs) were first defined as DNA segments of one kilobases (kb) or more that are present at variable copy numbers when compared to the reference genome (Feuk, et al. 2006). In recent times the description of CNVs has widened to include genomic changes as short as 50 bases (Alkan, et al. 2011). CNVs can be broadly classified into three classes: amplifications, deletions and insertions (Figure 10). Amplifications are described as gains in copy number of a genomic region or chromosome (Figure 10). They can be tandem when present in adjacent regions or interspersed when separated or distributed along the same chromosome (intrachromosomal) or distinct chromosomes (interchromosomal). Deletions are defined as loss of gene copies (Figure 10). Deletions can be heterozygous (loss of one allele) or homozygous (loss of both alleles). Insertions are integration of new genomic DNA (Figure 10).

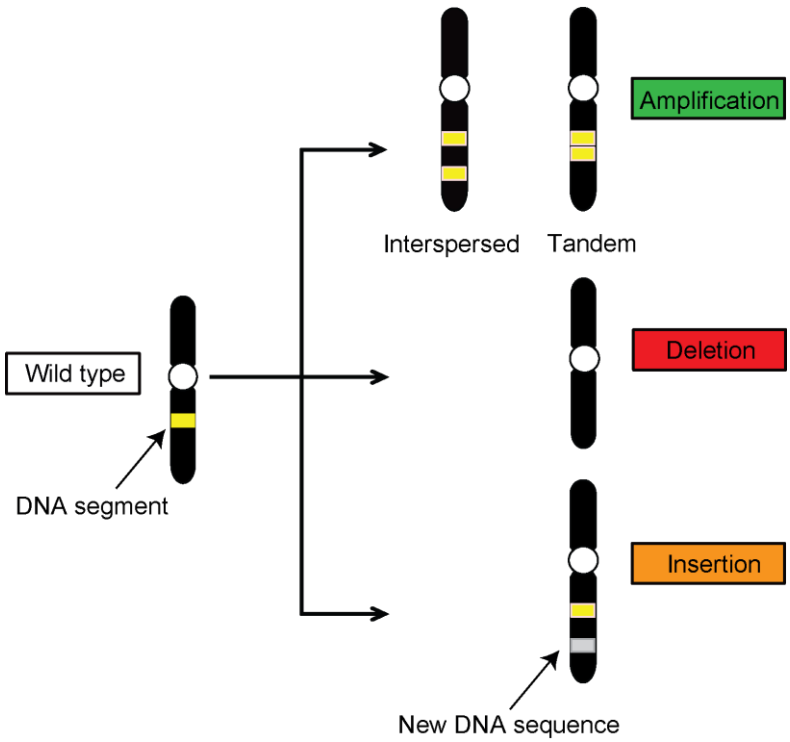


Figure 10: Classes of copy number variations

Shown are the three classes of CNVs. A gene or region of the genome (DNA segment) can undergo amplifications (copy number gains) or deletions (copy number losses) or may have insertion events. Amplifications can be interspersed or tandem duplications.

So far, four mechanisms have been reported (non-allelic homologous recombination (NAHR), non-homologous end joining (NHEJ), fork stalling and template switching (FoSTeS), and L1-mediated retrotransposition) for the formation of the majority of CNVs (Zhang, et al. 2009) (Figure 11, 12, 13, 14). NAHR is homologous recombination event induced by DNA double strand breaks (DSBs) and uses incorrect template for DNA repair (Figure 11). Studies have shown that NAHR preferentially occurs at hotspots inside low-copy repeats (LCRs) (Lupski 2004) and the segmental duplications caused by NAHR are between two LCRs (Stankiewicz and Lupski 2002). High sequence identity of non-allelic regions and subsequent crossover in the same chromosome may lead to tandem duplications and/or deletions when present in direct orientation (Figure 11) (Stankiewicz and Lupski 2002; Gu, et al. 2008; Hastings, Lupski, et al. 2009). A crossover between direct and inverted repeats in the same chromosome leads to deletions and inversions. Instead, crossovers between different chromosomes can result in chromosomal translocations (Stankiewicz and Lupski 2002; Gu, et al. 2008; Hastings, Lupski, et al. 2009). Although restricted to late S or G_2 (Lieber 2008), NAHR can occur in both meiotic as well as mitotic cells (Gu, et al. 2008). Homologous recombination events involving crossing over between homologous chromosomes can lead to loss of heterozygosity (LOH) if the same alleles segregate together at mitosis (Hastings, Lupski, et al. 2009).

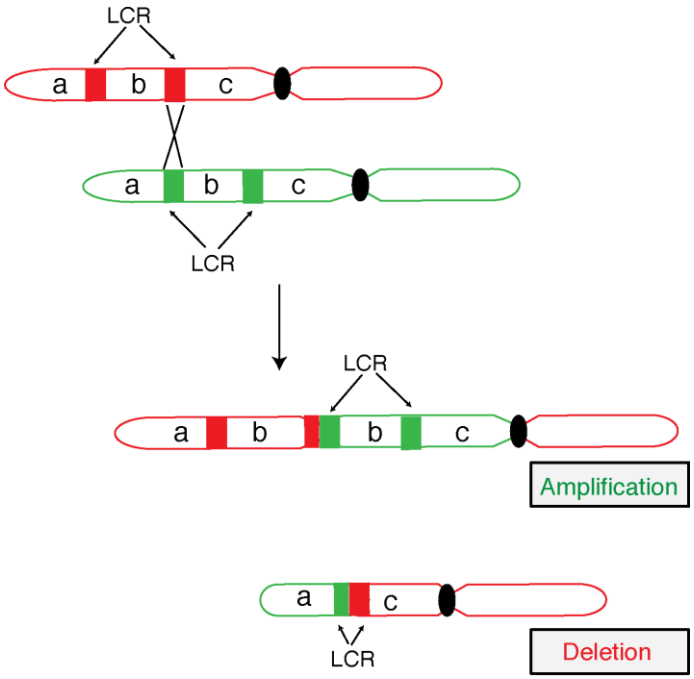


Figure 11: Mechanism of non-allelic homologous recombination (NAHR)

Shown is the mechanism of NAHR that occurs by unequal crossing over between flanking segmental duplications (represented by two red and two green bars on respective homologous chromosomes) resulting in reciprocal deletion and duplication of intervening sequence. LCR= low-copy repeat. Taken from (Malhotra and Sebat 2012).

NHEJ is one of the main mechanisms to repair DSBs in eukaryotic cells. It proceeds in four steps: i) detection of DSB; ii) bridging of the broken DNA ends; iii) modification of DNA ends to make them compatible for ligation; and iv) ligation of the DNA ends (Gu, et al. 2008) (Figure 12). NHEJ goes through many rounds of enzymatic activity. DNA ends are first recognized by the Ku protein and followed by recruitment of DNA-dependent protein kinase (DNA-PKcs). DNA-PKcs then mediate the activation of the Artemis nuclease, which trims back overhangs in preparation for ligation. Finally, ligation of the DNA ends is done by DNA ligase IV. Unlike NAHR, NHEJ does not require homology and rejoins DNA ends accurately or with small deletions. Alternatively it can also cause insertion of free DNA (Hastings, Lupski, et al. 2009). NHEJ can occur

throughout the cell cycle (Lieber 2008). Microhomology mediated end-joining (MMEJ) also known as alternative non-homologous end-joining, repairs DNA breaks via the use of microhomology and always results in deletions (McVey and Lee 2008; Hastings, Ira, et al. 2009) (Figure 12). MMEJ uses 5–25 base pair (bp) microhomologous sequences during the alignment of broken ends before joining, thereby resulting in deletions flanking the original break (McVey and Lee 2008). Unlike NHEJ, MMEJ uses Ku80 as the DNA binding protein (Hastings, Lupski, et al. 2009).

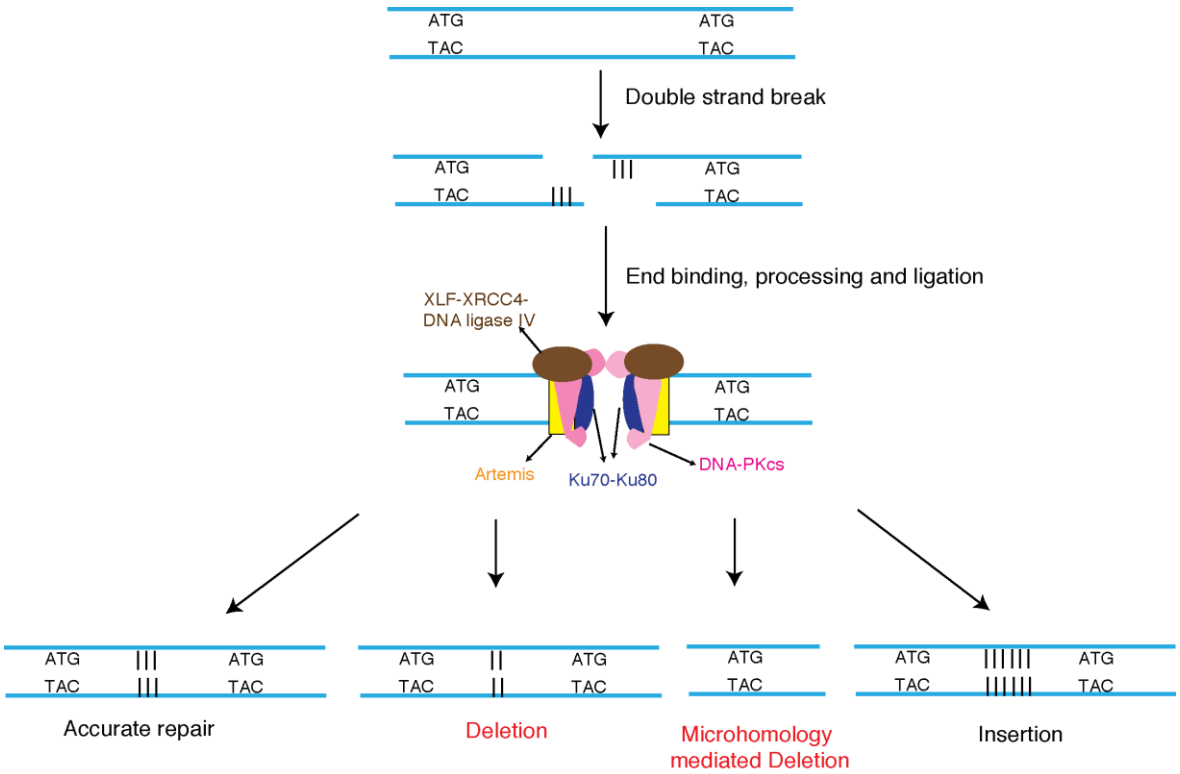


Figure 12: Mechanism of non-homologous end joining (NHEJ)

Shown is the mechanism of DNA repair by NHEJ. The different types of DNA double-strand breaks fixed by NHEJ combined with other alternate repair mechanisms, including microhomology mediated end-joining (MMEJ), leads to diverse repaired products. Taken from (Malhotra and Sebat 2012) and (Downs, et al. 2007).

FoSTeS has been reported to explain nonrecurrent and complex genomic rearrangements (Lee, et al. 2007). According to this DNA replication model (Figure 13),

when a replication fork encounters a nick in a template strand, one arm of the fork breaks off, thus producing a collapsed fork. At the single double-strand end, the 5′ end of the lagging strand is resected, giving a 3′ overhang. The 3′ single-strand end of lagging-strand template invades the second fork guided by regions of microhomology, forming a new low-processivity replication fork (Gu, et al. 2008). The extended end may dissociate and invade new forks repeatedly to form complex rearrangements (Lee, et al. 2007). Template switch occurring in front of or behind the position of the original collapse leads to deletion or duplication respectively (Malhotra and Sebat 2012). Microhomology-mediated break-induced replication (MMBIR) is a further generalized model of FoSTeS. MMBIR is used to repair DNA breaks when stretches of single-stranded DNA are available and share microhomology with the 3′ single-strand end from the collapsed fork (Hastings, Ira, et al. 2009).

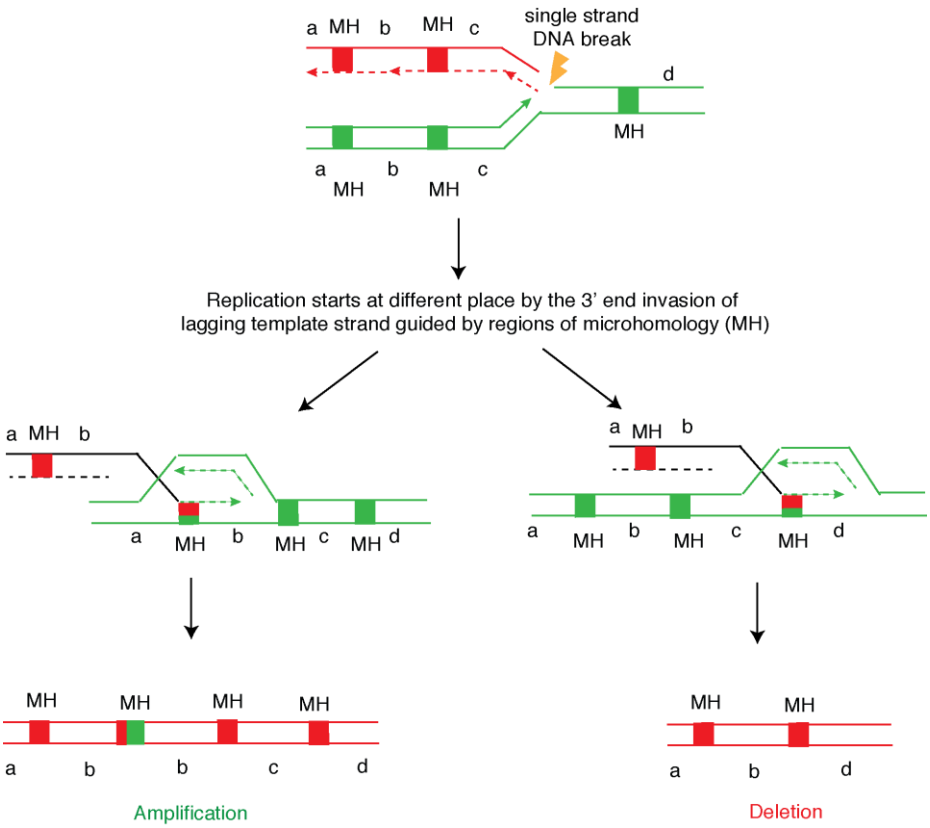


Figure 13: Mechanism of fork stalling and template switching (FoSTeS)

Shown is the mechanism of FoSTeS. FoSTeS occurs as a consequence of single strand breaks and invasion to other template strands for DNA synthesis resulting in complex genomic rearrangements. Taken from (Malhotra and Sebat 2012).

L1 are the only autonomous transposable elements in the human genome. L1-mediated retrotransposition causes rearrangements in the genome via RNA-mediated mechanisms. Though the mechanism is not clearly understood, it has been proposed to occur in five steps (Cordaux and Batzer 2009). In the first step, genomic L1 is transcribed by RNA polymerase II from an internal promoter that directs transcription initiation at the 5′ boundary of the L1 element. The L1 RNA is next exported to the cytoplasm where ORF1 (encoding an RNA-binding protein) and ORF2 (encoding a protein with endonuclease and reverse-transcriptase activities) are translated. Both proteins exhibit strong *cis*-preference and preferentially associate with the L1 RNA transcript that encoded them, to produce a ribonucleoprotein (RNP) particle. The RNP is then transported back into the nucleus by a mechanism that is poorly understood. L1 element is then integrated into the genome mostly likely by target-primed reverse transcription (Figure 14). Hallmarks of the integration process include frequent 5′ truncations, presence of an oligo dA-rich tail at the 3′end, and 2-20 bp-long duplications of the target site (Cordaux and Batzer 2009).

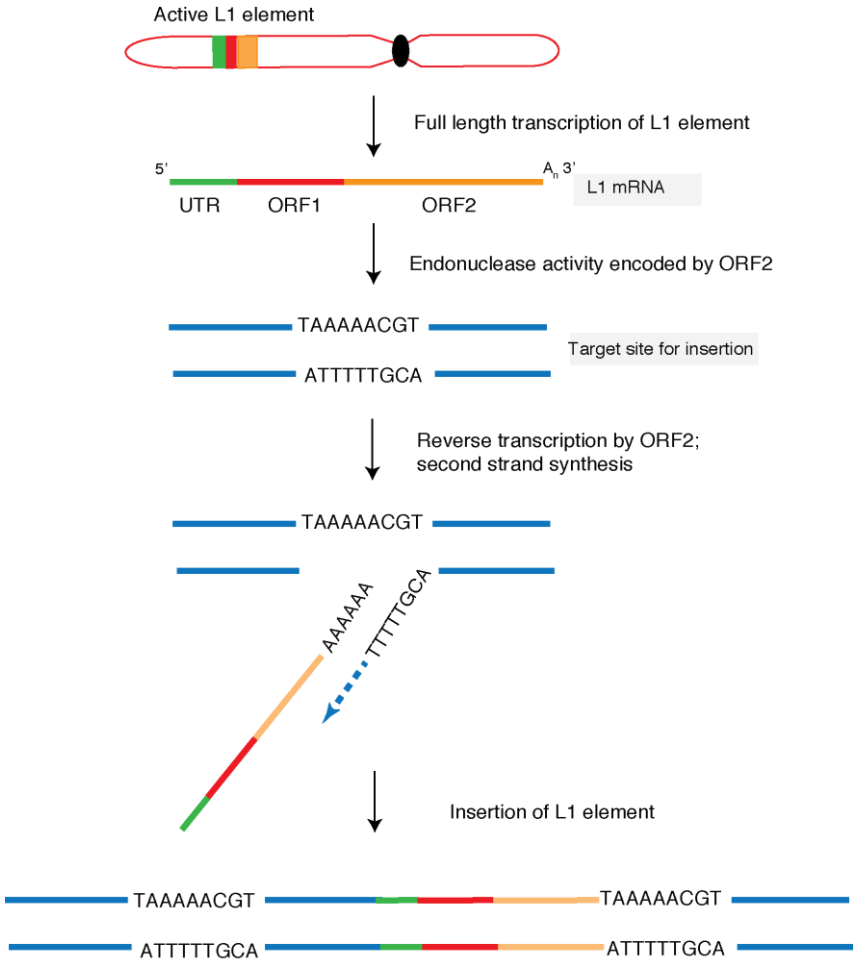


Figure 14: Mechanism of L1-reterotransposition mediated DNA repair

Shown is the mechanism of L1-reterotransposition mediated DNA repair. L1 elements are transcribed and inserted into the DNA target site. Taken from (Malhotra and Sebat 2012). ORF=open reading frame

1.6.2 Copy number variations in evolution and diseases

CNVs have been suggested to play a major driving role in evolution (Redon, et al. 2006; Zhang, et al. 2009). Gene duplication was proposed as the easiest way to produce new genes as early as 1970 by Susumu Ohno (Ohno 1970; Wolfe 2001). CNVs account for ~12% of variability in human genome (Redon, et al. 2006). CNVs that reach a population frequency of greater than 1% are referred to as copy number polymorphisms (Redon, et al. 2006). CNVs contain hundreds of genes and other functional elements. Significant

relationship between the genomic regions affected by CNVs and gene content has been reported (Cooper, et al. 2007). For instance regions with high gene density are enriched for CNVs (Cooper, et al. 2007). Additionally, CNVs in humans have been reported as enriched for olfactory and immunity genes as well as in genes encoding secreted proteins (Nguyen, et al. 2006). Furthermore, comparable studies of CNVs between humans and chimpanzees have identified functional categories of genes that are likely fixed by positive selection and involved in the adaptive phenotypic differentiation between the two species. These genes are particularly involved in inflammatory response and cell proliferation (Perry, et al. 2008). Moreover, selective advantages of CNVs in genes such as the salivary amylase gene, *AMY1,* the chemokine *CCL3L1* and the α-globin have been reported in humans (Higgs, et al. 1989; Gonzalez, et al. 2005; Perry, et al. 2007). The average copy number of *AMY1* in populations with high intake of starch is higher in comparison to populations with low intake (Perry, et al. 2007). High copy number of *CCL3L1* has been proposed to reduce the risk of HIV infection (Gonzalez, et al. 2005) while heterozygous deletion of α-globin confers resistance to malaria (Higgs, et al. 1989).

CNVs can alter transcription of genes by altering dosage, disrupting proximal or distant regulatory regions or through perturbation of transcript structure (Henrichsen, Chaignat, et al. 2009) (Figure 15). For example, genes undergoing amplifications or deletions may be over-expressed or under-expressed, respectively (Figure 15). A weak, yet positive correlation between CNVs and gene expression has been reported in mouse and rats (Guryev, et al. 2008; Henrichsen, Vinckenbosch, et al. 2009). CNVs have also been reported to influence the expression of genes located in their vicinity (Guryev, et al. 2008; Henrichsen, Vinckenbosch, et al. 2009). For example, deletions affecting silencers or insulator elements may result in increased gene expression of the transcript (Weischenfeldt, et al. 2013). Furthermore, CNVs have been identified as a key contributor to natural phenotypic variations (Cahan, et al. 2009).

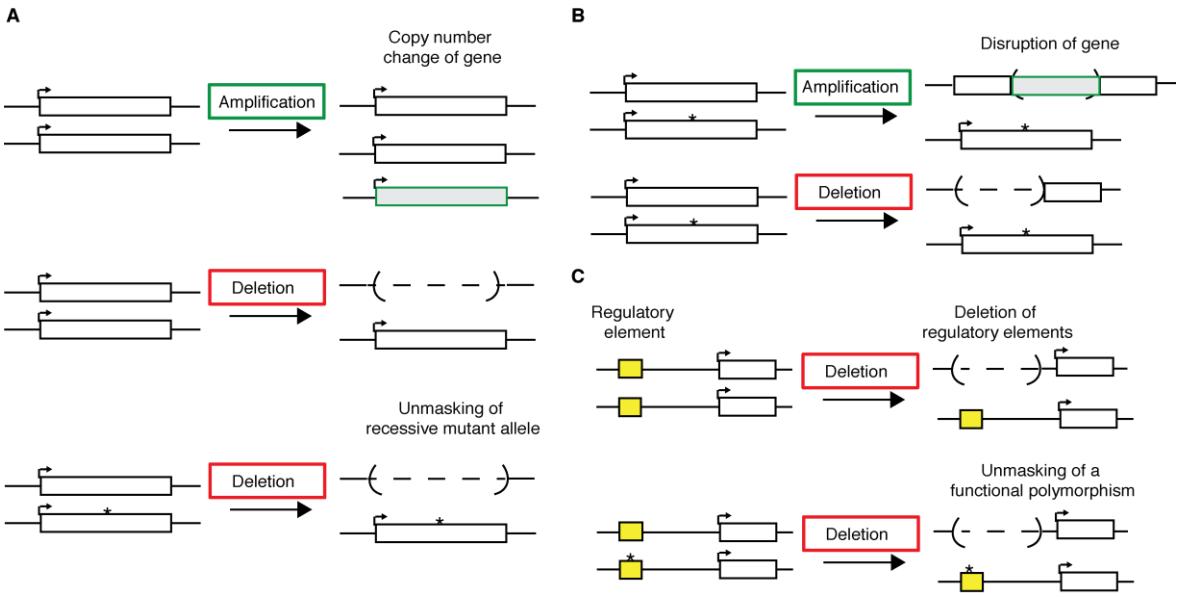


Figure 15: Effect of CNVs on expression

Shown are the possible ways in which expression of genes may be influenced due to CNVs. A) Amplification or deletion of gene may result in over-expression or under-expression. Gene expression can also be affected by a deletion of the allele that masks a recessive mutation. B) CNVs may disrupt genes either by deletion or amplification, resulting in reduced expression of genes. C) Expression of genes can also be altered by the deletion of important regulatory elements or of the allele masking a functional polymorphism within an effector. Taken from (Feuk, et al. 2006).

CNVs have been associated with many genetic disorders and complex diseases, including glomerulonephritis, autism, and schizophrenia (Table 3). For example, deletion of Fc fragment of IgG, low affinity IIIb, receptor (CD16b) (*FCGR3B)* has been associated with glomerulonephritis (Aitman, et al. 2006). Amplification of protease, serine, 1 (trypsin 1) (*PRSS1*) has been associated with the onset of hereditary pancreatitis (Le Marechal, et al. 2006).

Table 3: CNVs associated with genetic and complex diseases

Disease	*Gene / Locus*	*CNV*	*Reference*

Disease	Gene / Locus	CNV	Reference
Alzheimer	*APP*	Amplification	(Rovelet-Lecrux, et al. 2006)
Autism	*16p11.2*	Amplification / deletion	(Weiss, et al. 2008)
Autoimmunity	*FCGR3B*	Deletion	(Fanciulli, et al. 2007)
Body mass index	*NEGR1*	Deletion	(Willer, et al. 2009)
Charcot–Marie–Tooth neuropathy type 1	*17p11.2*	Amplification	(Raeymaekers, et al. 1991)
Crohn's disease of the colon	*DEFB4*	Deletion	(Fellermann, et al. 2006)
Crohn's disease	*IRGM*	Deletion	(McCarroll, et al. 2008)
Epilepsy	*15q13.3*	Deletion	(Helbig, et al. 2009)
Glomerulonephritis	*FCGR3B*	Deletion	(Aitman, et al. 2006)
Hereditary pancreatitis	*PRSS1*	Amplification	(Le Marechal, et al. 2006)
IMR	*15q13.3*	Deletion	(Mefford, et al. 2008; Sharp, et al. 2008)
IMR	*1q21.1*	Deletion	(Mefford, et al. 2008)
Parkinson's	*SNCA*	Amplification	(Singleton, et al. 2003)
Psoriasis	*LCE3C*	Deletion	(de Cid, et al. 2009)
Schizophrenia	*15q13.3*	Deletion	(International Schizophrenia 2008; Stefansson, et al. 2008)
Schizophrenia	*1q21.1*	Deletion	(International Schizophrenia 2008; Stefansson, et al. 2008)
Schizophrenia	*22q11.2*	Deletion	(Bassett, et al. 2008)
Susceptibility to HIV-1	*CCL3L1*	Deletion	(Gonzalez, et al. 2005)
Systemic lupus erythematosus	*C4*	Deletion	(Yang, et al. 2007)

Reported are the diseases, genes/locus, type of CNVs and the reference.

1.6.3 Copy number variations and cancer

Genomic instability has been identified as the most prominent feature of cancer cells that may orchestrate the acquisition of cancer hallmark capabilities (Hanahan and Weinberg 2011). Somatically acquired amplifications, deletions, and copy neutral loss of

heterozygosity (CN-LOH) events frequently occur in cancer (Ciriello, et al. 2013; Zack, et al. 2013). These genomic copy number alterations may occur in a small genomic region of the chromosome (focal alterations), or may affect whole chromosome (aneuploidy) or in extreme situations may lead to chromosomal instability (CIN) involving abnormal copy number states of many chromosomes in the genome (Figure 16) (Tang and Amon 2013).

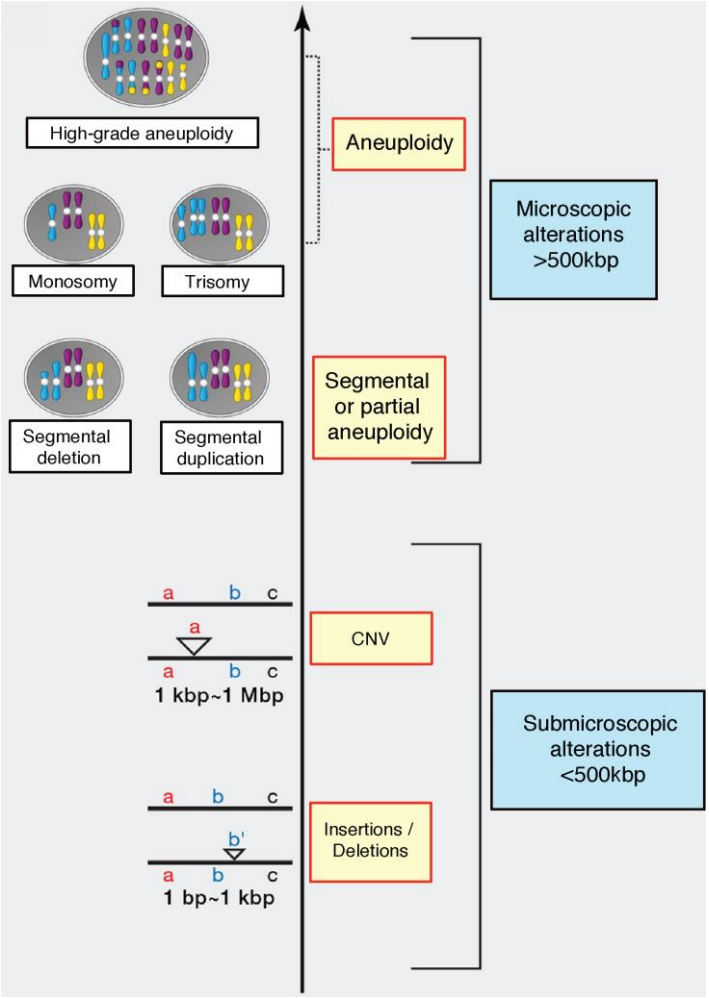


Figure 16: Genomic copy number alterations in cancer

Shown are the microscopic (aneuploidy, CIN) and submicroscopic (focal) genomic alterations that are observed cancers. Taken from (Tang and Amon 2013).

Somatic CNVs may play passenger or driver roles in tumorigenesis. For instance, CIN and aneuploidy observed in retinoblastoma are a consequence of the inactivation of retinoblastoma-associated protein and they play a passenger role in cancer development

(Zheng and Lee 2002; Gordon, et al. 2012). However, many recurrent CNVs have been identified as pathogenic events in solid tumours. For instance, focal amplification of the oncogene human epidermal growth factor receptor 2 (*HER2*) is observed in 30% of breast cancers (Slamon, et al. 1987; Cameron, et al. 2008). Similarly, focal amplification of the proto-oncogene *N-myc*, occurs in ~30% of advanced neuroblastomas and is associated with rapid tumour progression (Seeger, et al. 1985). Deletion of the tumour suppressor gene phosphatase and tensin homolog (*PTEN*) is observed in 68% of primary prostrate cancer (Trotman, et al. 2003; Yoshimoto, et al. 2006). Driver CNVs, tend to reside in the same genomic regions across different cancers, especially focal modifications (Beroukhim, et al. 2010; Kim, et al. 2013; Zack, et al. 2013). High-resolution studies of somatic CNVs across multiple cancer types have identified prevalence of arm-level alterations (Beroukhim, et al. 2010; Kim, et al. 2013; Zack, et al. 2013). Whole-chromosome alterations are recurrently observed in several cancer types. For example, gain of chromosome 8 is seen in 10–20% of acute myeloid leukaemia, as well as in some solid tumours, including Ewing's Sarcoma and desmoid tumours (Qi, et al. 1996; Maurici, et al. 1998; Paulsson and Johansson 2007). Trisomy of chromosome 7 harbouring non-random duplication of mutant *MET* allele has been implicated in hereditary papillary renal carcinoma (Zhuang, et al. 1998).

In addition to somatically acquired CNVs, inherited germline CNVs have also been associated with cancer and risk for cancer (Table 4). For example, an inherited CNV of *NBPF23* is associated with the onset of neuroblastoma (Diskin, et al. 2009). Similarly, deletion at chromosome 2p24.3 was observed to confer stronger risk of aggressive prostate cancer (Liu, et al. 2009). Other examples include rare 4q13 duplication in melanoma-predisposed family, germline microdeletion of 9p21.3 containing *KIA1797* and *MIR491* genes in colorectal (Venkatachalam, et al. 2011) and breast cancers (Krepischi, Achatz, et al. 2012), respectively.

Table 4: Literature review of whole-genome studies associating germline CNVs with cancer susceptibility.

Type of cancer	*Patients*	*CNVs*	*Germline CNV details*	*Reference*
***Familial pancreatic cancer**†*	57	56	Rare CNVs (not present in the DGV and 607 study controls)	(Lucito, et al. 2007)
***Familial and early-onset colorectal cancer**†*	41	7	Rare CNVs (not present in the DGV and 1600 controls from in-house database)	(Venkatachalam, et al. 2011)
Familial and early-onset breast cancer*	68	26	Rare CNVs (not present in the DGV, 100 study controls and 158 from in-house database)	(Krepischi, Achatz, et al. 2012)
***Familial melanoma**†*	30	1	Rare 4q13 duplication (three affected individuals from the same family; CXC genes)	(Yang, et al. 2012)
Aggressive prostate cancer	498	1	Recurrent 2p24.3 deletion (12.63% in patients in comparison to 8.28% in 494 study controls)	(Liu, et al. 2009)
Neuroendocrine tumours of the ileum	226	4	Recurrent 18q22.1 deletion (no known genes; (6.19% in patients in comparison to 2.06% in 97 study controls)	(Walsh, et al. 2011)
Neuroblastoma	846	1	Recurrent CNV at 1q21.1 (gene *NBPF23*; odds ratio 2.49 [803 study controls])	(Diskin, et al. 2009)
Hepatocellular carcinoma	386	6	Recurrent CNV at 1p36.33 (odds ratio 17.0 [687 study controls])	(Clifford, et al. 2010)
Nasopharyngeal carcinoma	278	8	Recurrent 6p21.33 deletion (*MICA* and *HCP5* genes) Gender-specific association (male)	(Tse, et al. 2011)

Reported are the type of cancer cohort, number of unrelated patients, relevant CNVs, details on the germline CNVs identified and references of the respective studies. Source (Krepischi, Pearson, et al. 2012). †Studies of familial cancer or high-risk cancer patients, DGV=Database of Genomic Variants.

CNVs may contribute to cancer development by altering the expression of tumour suppressors, oncogenes or non-coding RNAs such as miRNAs. For instance, amplification of oncogene *MAPKAPK2* shows increased mRNA and protein expression and is associated with an increased risk and poor prognosis for lung cancer in Chinese populations (Liu, et al. 2012). *RUNX3*, a tumour suppressor in gastric cancer is not expressed in 45%-60% of

human gastric cancer cells of which around one-third are due to hemizygous deletions (Li, et al. 2002). In addition, fusion gene *TMPRSS2-ERG* formed due to interstitial deletion in chromosome 21 is frequently observed in prostrate cancer (Tomlins, et al. 2005). CNVs reported in database of genomic variants disrupt nearly 40% of cancer-related genes (Shlien and Malkin 2010). Though a linear correlation between CNVs, mRNA and protein expressions have not been observed in cancers (Geiger, et al. 2010; Zhang, et al. 2014), proteins encoded by amplified oncogenes are often overexpressed (Zhang, et al. 2014).

1.6.4 Methods for detecting copy number variations

Since CNVs may play an important role in the development of many diseases including cancer, their identification is essential for diagnosis and treatment. There are three approaches for detecting CNVs: (i) hybridization-based microarrays; (ii) single-molecule analysis; and (iii) next generation sequencing (NGS) (Alkan, et al. 2011).

Hybridization-based microarrays are represented by array comparative genomic hybridization (aCGH) and single nucleotide polymorphism (SNP) arrays. Microarrays are composed of in situ hybridization of fluorescently labelled genomic DNA regions corresponding to short or long oligonucleotides (probes). CNVs from microarrays are then detected based on the intensity of the fluorescence. In case of aCGH, fluorescently labelled genomic DNA from test and reference samples are hybridized to target probes. CNVs are then detected based on the signal ratio. In SNP arrays, the genomic DNA of test sample is hybridized to short probes with single nucleotide difference. These arrays are used for both genotyping and copy number analysis. Though microarrays are cost effective, they are limited to identifying CNVs in reference genome that are used for designing the probes (Alkan, et al. 2011). Furthermore, microarrays are less sensitive in detecting single copy gains compared to copy loss, thus suggesting bias in detecting CNVs (Greshock, et al. 2007).

Single-molecule analysis includes techniques such as FISH, spectral karyotyping and optical mapping that identifies the structure and location of CNVs. These methods are particularly useful in identifying balanced structural genomic rearrangements that cannot be identified using microarrays. Although these methods are capable to detect novel insertions, they have limited throughput and low resolution (Alkan, et al. 2011).

The advent of NGS has heavily influenced structural variation studies. NGS techniques are used for whole genome sequencing (WGS), whole exome sequencing (WES) or targeted re-sequencing screenings. NGS allows CNV detection at breakpoint resolution and the estimation of absolute copy numbers. Most of the current algorithms use read-pair or read depth approach for detecting CNVs (Figure 17). Read-pair methods assess the distance between paired-end reads and their orientation in comparison to the reference genome to identify CNVs at breakpoint (Tuzun, et al. 2005) (Figure 17). Deletions are defined by read-pairs that map far from each other, whereas insertions are described by read-pairs mapping very close to one another. Tandem duplications are identified by read-pairs that are inconsistent in their orientation (Figure 17). Read-pair methods are not suitable for identifying CNVs from WES data because CNVs that have breakpoints in untargeted regions cannot be detected. Read depth is another approach for detecting CNVs from NGS. This approach assumes that a duplicated region will have higher number of reads compared to a diploid genome, whereas a deleted region will have lower number of reads (Campbell, et al. 2008) (Figure 17). Read depth approaches can be applied for detecting CNVs from both WGS and WES screenings.

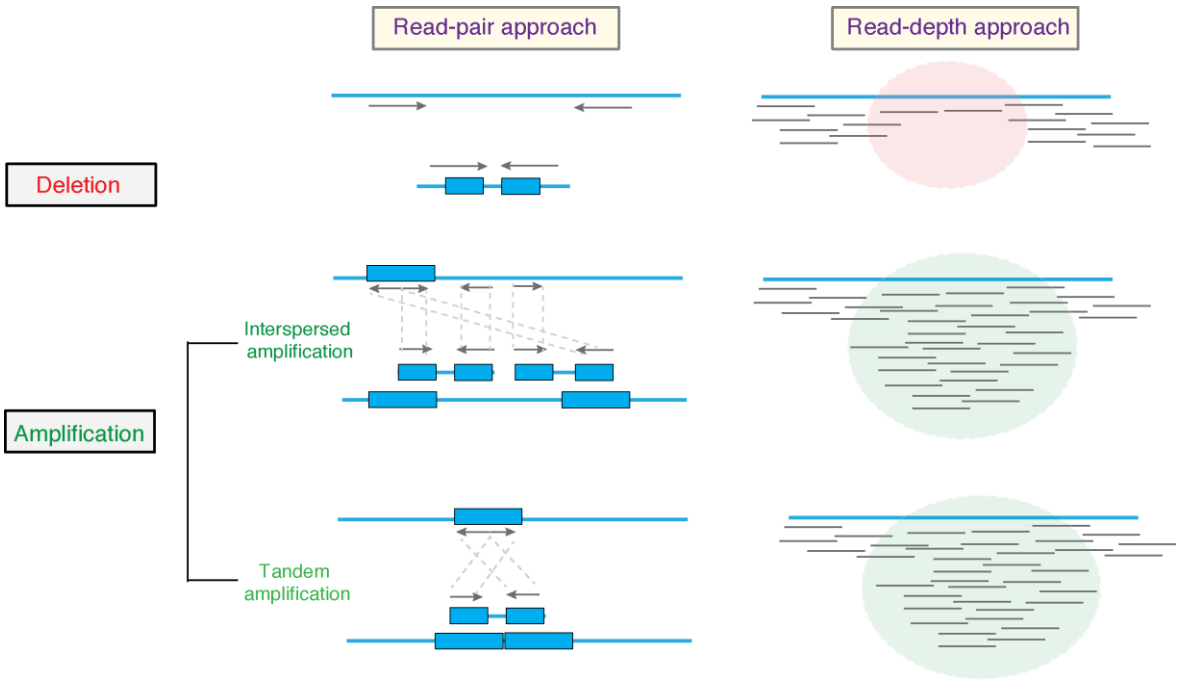


Figure 17: Approaches for CNV detection from next generation sequencing data

Shown are the two approaches for detecting CNVs from NGS. Read-pair approach is more suitable for WGS data, whereas read depth approach can be applied to both WGS and WES screenings. Taken from (Alkan, et al. 2011)

Although WGS provides the most comprehensive CNV profile, WES remains the most widely used NGS approach because it is still time and cost effective. In addition, WES gives insights into the genomic alterations of protein coding genes, which are often the most interesting to follow up. Detection of CNVs from WES data is however challenging due to the different size and sequence composition of exons, which result in non-uniform sequence coverage (Hodges, et al. 2007; Magi, et al. 2012; Sims, et al. 2014) (Figure 18). In addition, since exons are located at variable distances from each other, the detection of CNVs at breakpoint resolution is challenging (Liu, et al. 2013; Zhao, et al. 2013) (Figure 18). Several methods have been developed in recent years to detect CNVs from WES data including ExomeCNV (Sathirapongsasuti, et al. 2011), VarScan2 (Koboldt, et al. 2012), ControlFreeC (Boeva, et al. 2012), ADTEx (Amarasinghe, et al. 2013), and EXCAVATOR (Magi, et al. 2013).

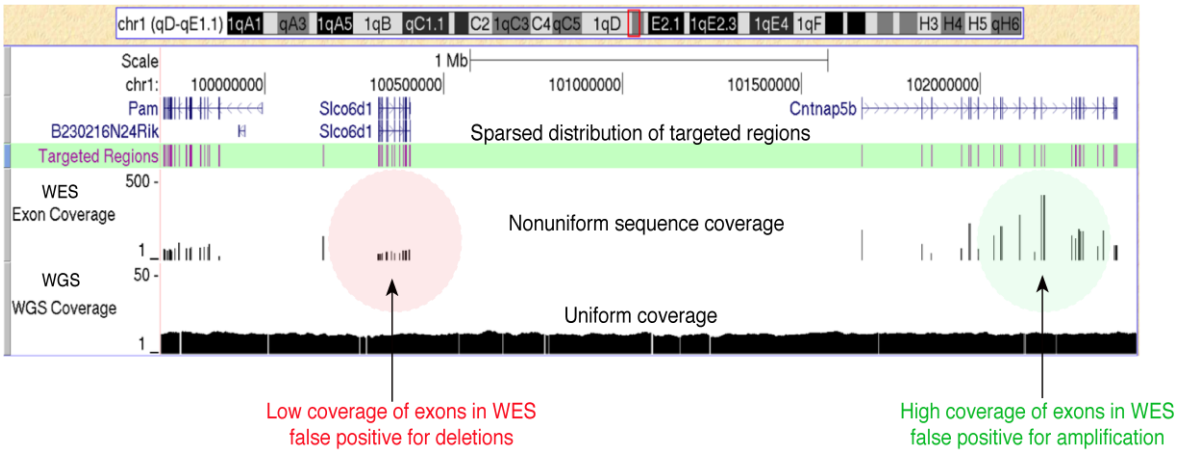


Figure 18: Pictorial representation of challenges in detecting CNVs from WES data

Shown is the read depth and genome coverage in WES data and WGS screenings. CNV detection from WES data is affected due to bias in exon coverage that lead to non-uniform coverage resulting in false discovery of CNVs.

ExomeCNV is the first method that was used for detecting CNVs from WES data. It compares the read depth of exons between tumour and normal samples and employs circular binary segmentation (CBS) to identify breakpoints of copy number change in the tumour. ExomeCNV is prone to ambiguous calls because it does not normalizes for the difference in the coverage between the samples. VarScan 2 is another widely used method that normalizes the coverage between tumour and normal samples and performs segmentation using CBS to identify genomic segments of copy number change. It is unable to identify CN-LOH events. EXCAVATOR is a recently published method that performs three steps of normalization to remove sources of variations due to GC content and presence of repetitive sequences. It uses a new segmentation algorithm that exploits the distance between the adjacent exons, thus accounting for the sparseness of the exons in WES. Similar to VarScan 2, EXCAVATOR does not detect CN-LOH events.

Although using different technical solutions, the majority of these methods identify CNVs between two or more samples (e.g. tumour and matched normal) after dividing the genome sequence into segments. Segmentation works by merging adjacent genomic regions into longer segments in the attempt to minimize coverage variability within the

segments while maximizing it between them. Segmentation has been widely applied to aCGH (Tonon, et al. 2005; Kabbarah, et al. 2010) and WGS (Campbell, et al. 2008; Zhang, et al. 2013) data, where information is available for long and continuous regions. Although it has been adapted by exome-based methods to detect CNVs, the scattered nature of the data and the high variability in exon coverage may reduce its efficacy in this context. Additionally, these methods use normalizations that rely on the assumption that CNVs do not affect many genes. However, this is not applicable to cancers that may undergo large-scale rearrangements. These exome-based methods do not correct for this assumption while calling CNVs, which may reflect in identification of ambiguous CNVs.

1.7 Aim of the thesis

The aim of my PhD project is to identify the genomic alterations that occur during liver cancer progression in the absence of external mutagens

HCC arises in response to mutagens such as virus infection, aflatoxin and alcohol, or as a consequence of metabolic diseases including obesity and diabetes (Block, et al. 2003; El-Serag and Rudolph 2007). As a result, HCC is heterogeneous in terms of genetic make up (Unsal, et al. 1994; Dragani 2010). Recent studies showed a strong dependence of the acquired mutation signature on the underlying mutagenesis mechanism, thus suggesting that the genetic heterogeneity in HCCs may depend on the causative agents (Zhang 2012). HCC is almost invariably associated with an underlying inflammatory state regardless of the initiating agent. However, the contribution of chronic inflammation to the acquisition of cancer driver genomic changes is still unclear.

To understand the molecular basis for tumorigenesis and progression in liver cancer induced due to chronic inflammation, we studied the genomic landscape of human BSEP-deficient HCCs (BSEP-HCCs). These tumours develop in response to chronic liver injury due to inflammation resulting from accumulation of bile salts (Knisely, et al. 2006) in the

absence of external mutagens. To further investigate the role of alterations that occur in BSEP-HCCs, we sequenced tumours from *Mdr2*-KO mice. *Mdr2*-KO is a genetic mouse model that has similar etiopathogenesis as that of BSEP-HCCs. Thus, *Mdr2*-KO mouse model offers an ideal system for studying the cancer progression in these types of tumours.

We mapped the genomic alterations in both human and mouse HCCs. We used known methods to detect mutations and indels from exome sequencing data and CNVs from SNP arrays and whole genome sequencing data. In addition, we developed a novel method, GeneCNV, to identify CNVs from whole exome and targeted re-sequencing screenings. We next characterized the CNV spectrum in both human and mouse HCCs and observed a consistent genomic signature within and between species that was unique to these types of tumours.

Methods

2.1 Sample description

2.1.1 Samples from human liver cancer

Seven children were diagnosed with BSEP-HCC (Table 5). The background liver in all patients exhibited parenchymal rather than portal-tract cholestasis, with bile salt export pump expression detectable in none of them. Some patients had frank cirrhosis and others had only fibrosis, which varied in degree from patient to patient (Table 5). Samples 7860 175, 1790, 2896, and UKT came from single unencapsulated masses, while sample 23836 was derived from one of the several HCCs within a single liver. Sample HB4R was a relapse that developed within allograft liver 6 years after transplantation. This patient underwent chemotherapy before relapse and chirurgical resection. All patients had mutations in ABCB11. (Table 5)

Table 5: Description human BSEP-HCC samples

ID	*Storage*	*Tumour content*	*Background liver**	*Sex*	*Age*	*ABCB11 non-silent SNVs and indels*	*Zygosity*	*Modification(s) on protein*	*HGMD*	*MutPred deleterious effect*	*BSEP expression in the tumour*	*BSEP expression in background liver*	*Histopathology findings*
175	Frozen	90%	Extensive fibrosis	M	1Y 6M	937C>A	Heterozygous	Arg313Ser	DM	Medium	NO	NO	Two tumours; both trabecular
						1331T>C	Heterozygous	Val444Ala	DFP	Very low			
						1445A>G	Heterozygous	Asp482Gly	DM	Very high			
						2316T>A	Heterozygous	Tyr772X	DM	-			
7860	FFPE	90%	Mild fibrosis	F	2Y 6M	1331T>C	Heterozygous	Val444Ala	DFP	Very low	NO	NO	Multiple tumours (>10); trabecular, pseudoglandular and clear-cell
						2429delT	Heterozygous	Leu810PhefsX11	-	-			
23836	FFPE	90%	Cirrhosis	M	1Y 3M	1331T>C	Heterozygous	Val444Ala	DFP	Very low	NO	NO	Single tumour; trabecular
						1445A>G	Heterozygous	Asp482Gly	DM	Very high			
						1462T>C	Heterozygous	Ser488Pro	-	-			
						1628A>C	Heterozygous	Asp543Ala	-	-			
HB4R	Frozen	70%	Scarring; no cholestasis	F	8Y 6M	1331T>C	Homozygous	Val444Ala	DFP	Very low	NO	YES (allograft)	Single tumour; clear-cell

ID	*Storage*	*Tumour content*	*Background liver**	*Sex*	*Age*	*ABCB11 non-silent SNVs and indels*	*Zygosity*	*Modification(s) on protein*	*HGMD*	*MutPred deleterious effect*	*BSEP expression in the tumour*	*BSEP expression in background liver*	*Histopathology findings*
1790	Frozen	60%	Moderate fibrosis with bridging	M	11Y 7M	378del	Homozygous	Thr127HisfsX7	-	-	NO	NO	Single tumour; trabecular
2896	Frozen	50%	Cirrhosis	F	1Y 3M	1331T>C	Homozygous	Val444Ala	DFP	Very low	NO	NO	Three tumours; trabecular and clear-cell
						1416T>A	Homozygous	Tyr472X	DM	-			
						2029A>G	Homozygous	Met677Val	-	-			
						3556G>A	Homozygous	Glu1186Lys	FP	Low			
UKT	Frozen	40%	Cirrhosis	M	1Y 3M	1460G>C	Homozygous	Arg487Pro	DM	Medium	NO	NO	Single tumour; trabecular

For each lesion reported are details on the sample, the histopathological description. All patients had non-remitting cholestasis of neonatal onset (PFIC). BSEP expression was assessed immunohistochemically in tumours and corresponding background livers. FFPE = formalin-fixed paraffin embedded; *Cholestasis was observed, unless otherwise stated. DM=disease-causing mutations, DFP=disease-associated polymorphism with additional supporting functional evidence and FP= in vitro/laboratory or in vivo functional polymorphism

2.1.2 Samples from mouse liver cancer

Founders of the FVB.129P2-Abcb4tm1Bor/J (*Mdr2*-KO, stock number: 002539) and FVB/NJ (*Mdr2* wild type, stock number:001800) mice were purchased from The Jackson Laboratory and colonies of both strains maintained under specific pathogen-free conditions. Mice were sacrificed at 10-16 months (Table 6).

Table 6: Description of mouse samples

ID	*Targeted region*	*Tumour content*	*Tumour size (cm)*	*Age (months)*	*Sex*
51509/1	Whole exome	20%	1.1	16	M
60400/2	Whole exome	40%	1.4	13	F
218/1	Whole exome	50%	1.0	15	M
52686/1	Whole exome	50%	0.7	15	F
58853/3	Whole exome	60%	1.7	15	M
60400/1	Whole exome / whole genome	60%	0.6	13	F
58163/3	Whole exome	70%	3.0	15	M
58163/4	Whole exome	70%	3.0	15	M
215/1	Whole exome	80%	1.8	14	M
54913/10	866 genes	ND	0.1	10	F
54913/8	866 genes	ND	0.5	10	F
55481/10	866 genes	ND	0.3	10	F
55484/4	866 genes	30%	3.0	10	F
218/3	Whole genome	70%	1.0	15	M

2.2 Experimental procedure

We performed all experiments in collaboration with Giacchino Natoli's group, Department of Experimental Oncology, European Institute of Oncology (IEO), Milan, Italy. Agnese Collino maintained the mouse colonies and performed the experiments. Federica Pisati prepared the sample for histology inspection. Histologpatholgy inspection on mouse and human samples was done by Enrico Radaelli from VIB Center for the Biology of Disease, KU Leuven Center for Human Genetics and A.S.Kinsely from Institute of Liver Studies, King's College Hospital respectively. DNA sequencing, TaqMan copy number assay and gene expression experiments were performed by Consortium for

Genomic Technologies (Cogentech). SNP microarray experiments for human samples were conducted by Fondazione Filarete.

2.2.1 DNA extraction

For human BSEP-deficient HCCs, we obtained frozen or formalin-fixed paraffin-embedded (FFPE) samples from seven patients during native-liver hepatectomy with parental written consent. We used non-neoplastic liver tissues from all the patients as matching background references. We extracted genomic DNA from each tumour and matched background liver tissue using the DNeasy Blood and Tissue Kit (Qiagen) for frozen samples and with the AllPrep DNA/RNA FFPE Mini Kit (Qiagen) for FFPE blocks.

For mouse *Mdr2*-KO HCCs, we snap froze the adenoma and HCC nodules from *Mdr2*-KO mice for DNA/RNA extraction or fixed in formalin for histological analysis and performed the initial pathological stage (inflammation) DNA/RNA extraction on purified populations of hepatocytes obtained via collagenase liver perfusion, using a two-step protocol. We snap froze the normal livers or kidneys and used them as the reference. We homogenized all the frozen tissue samples with GentleMACS Dissociator (Miltenyi Biotec) before column extraction and extracted the genomic DNA using the DNeasy Blood and Tissue Kit (QIAGEN) according to the manufacturer's protocol.

Agnese Colino from Giacchino Natoli's group had performed the DNA extraction.

2.2.2 DNA-sequencing

2.2.2.1 Whole exome sequencing of human samples

We performed target capture on six human tumours and matched normal samples (Table 7) using the SureSelect XT Human All Exon V4 kit (Agilent) targeting 20,965 human genes, following the manufacturer's protocol with minor modifications. We excluded sample UKT from whole exome sequencing because of the low tumour content.

We sheared around 3 ug of genomic DNA using Adaptive Focused Acoustics technology (Covaris) and selected 200 bp fragments after library preparation with Illumina DNA Sample Prep Kit using the Agencourt AMPure PCR Purification system (Beckman Coulter). We further amplified selected fragments with 5 to 7 cycles of PCR and hybridized 500 ng of DNA with the bait library followed by paired-read cluster generation on the Cluster Station (Illumina). We sequenced libraries of each sample using one half-lane of Illumina HiSeq2000, with 76 bp or 101 bp paired-end protocol, except for the tumoral sample of patient 7860, where one entire lane was used due to high levels of DNA degradation (Table 7).

DNA sequencing unit in Consortium for Genomic Technologies (Cogentech) had performed the DNA sequencing.

Table 7: Whole exome sequencing setting of BSEP-HCCs

ID	*Sequencing setting*	*Samples per lane*
175	101 PE	2
7860	76 PE	1
23836	76 PE	2
HB4R	101 PE	2
1790	101 PE	2
2896	101 PE	2

For each lesion reported is the read length and number of lanes used for sequencing. All experiments were performed with a paired end setting.

2.2.2.2 Targeted sequencing, whole exome sequencing and whole genome sequencing of mouse samples

For targeted re-sequencing we designed SureSelect custom capture kit for mouse genes that are orthologs of known human cancer genes. In order to select genes for targeted re-sequencing, we first collected all human cancer genes from the Cancer Gene

Census (Futreal, et al. 2004), COSMIC (http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/) and high-throughput cancer mutational screenings (http://ncg.kcl.ac.uk/) that totalled to 2,061 genes. We next identified the mouse orthologs of these genes using eggNOG (Jensen, et al. 2008) and MGI (http://www.informatics.jax.org) that resulted in a total of 1,753 mouse orthologs, of which only 866 had RefSeq entries. We then designed the SureSelect Custom kit (Agilent) to capture 15,067 exons of these 866 genes, for a total of 2.7 Mbp of DNA. We also excluded exons that were shorter than 60 bp (except for those with mutations in COSMIC), sequence repeats, segmental duplications, PAR regions and gaps. Further, we selected regions with GC content ranging from 30 to 65% to optimize the capture efficiency.

We performed target capture using SureSelect custom kit for the 866 selected genes (~2.7 Mb) and the SureSelect XT Mouse All Exon kit (Agilent) targeting 21,543 mouse genes (~50.4 Mb) following the manufacturer's protocol with slight modifications (Table 8). Briefly, we sheared around 3 ug of genomic DNA using an ultrasonic disruptor (Bioruptor, Diagenode) or using Adaptive Focused Acoustics technology (Covaris) and selected 200-250 bp fragments after library preparation with the Illumina DNA Sample Prep Kit. We then purified the genomic fragments by gel extraction, or using the minelute PCR purification kit (QIAGEN), or using the Agencourt AMPure PCR Purification system (Beckman Coulter). We further amplified the fragments with 10 cycles of PCR and hybridized 500 ng of DNA with each bait library followed by single- or paired-read cluster generation on the Cluster Station (Illumina). We sequenced the libraries obtained for the 866 genes on the Genome Analyzer IIx with the 76 single end protocol, using one lane for each tumour sample or matching normal sample and the libraries obtained for the whole exomes using one-half lane of Illumina HiSeq2000 for each sample, with the 101 bp paired-end protocol (Table 8).

We performed whole genome sequencing of HCCs (218/3 and 60400/1) and matched normal samples from two *Mdr2*-KO mice: 218 and 60400. For whole genome

sequencing, we sheared around 1 ug of mouse genomic DNA in 400-500 bp fragments using an ultrasonic disruptor (Bioruptor, Diagenode) and prepared the libraries using Illumina Paired-End DNA Sample Prep Kit. We sequenced the obtained libraries using one lane of Illumina HiSeq2000 and 101 bp paired-end protocols (Table 8).

Fabio Iannelli from Francesca Ciccarelli's group created the custom microarray using Agilent eArray services and the DNA sequencing unit in Cogentech performed the DNA sequencing.

Table 8: Targeted and whole exome sequencing settings of *Mdr2*-KO HCC

ID	*Target regions*	*Targeted regions Mbps*	*Sequencing setting*	*Samples per lane*
51509/1	Whole exome	50.4	101 PE	2
60400/2				2
218/1				2
52686/1				2
58853/3				2
60400/1				2
58163/3				2
58163/4				2
215/1				2
54913/10	866 genes	2.7	76 SE	1
54913/8				1
55481/10				1
55484/4				1
218/3	Whole genome		101 PE	1
60400/1				1

For each lesion reported are details on read length, number of lanes used and the length of targeted genome used for capture and sequencing. All experiments were performed with a paired end setting.

2.2.3 Dilution experiment for assessments of variant calling

We measured the specificity of the variant calling procedure by re-sequencing 16 somatic non-silent SNVs in human tumours and 15 somatic non-silent SNVs mouse in mouse tumours with Sanger sequencing. For Sanger sequencing, we amplified genomic regions surrounding the somatic SNVs by PCR using the Taq DNA Polymerase (Qiagen)

and sequenced them in the tumour and corresponding reference in both directions on a 3730xl DNA Analyzer (Applied Biosystems) using the dRhodamine chemistry. Of the 16 SNVs in human, for two SNVs the PCR amplification failed and we confirmed 13 out of the 14 remaining SNVs (specificity = 92.9%). In mouse we also confirmed 14 out of the 15 SNVs (specificity = 93.3%).

In order to measure the sensitivity in calling somatic mutations in cancer samples with variable tumour contents due to contamination from normal tissue, we used eight dilutions of a homozygous germline mutation (CC) with the corresponding wild type genotype (TT). In this setting, the frequency of the variant allele mimics the decreased frequency of a somatic mutation in presence of normal tissue contamination.

We performed the dilutions of the minor allele as follows. We amplified a 105 bp long region centred on the mouse SNP rs32609672 (dbSNP build 128, chr5:36209358) from the genomic DNA of two mice with CC and TT homozygous genotypes (FVB/NJ and C57BL/6J strains, respectively), using a nested PCR approach. We first, amplified and Sanger sequenced a 393 bp fragment from each genomic DNA for genotype confirmation. Subsequently, we performed nested PCR to generate a 105 bp long amplicon, centred on the nucleotide of interest. We next purified the two amplicons (CC and TT) using the MinElute PCR Purification Kit (Qiagen) and used for library preparation with Illumina Paired-End DNA Sample Prep Kit. We then pooled the two libraries in eight different molar ratios with C:T proportion ranging from 0.04 to 0.41. These dilutions corresponded to variant allele (C) frequencies ranging from ~4% to ~41% and simulated a normal contamination ranging from ~20% to ~90%. We then sequenced each pool on a different Illumina GAIIx lane together with other samples in a ~1:1000 molar ratio and aligned the reads obtained to mouse chromosome 5 (NCBI37/mm9) using the same alignment setting, duplicate removal and somatic variant calling as described in paragraph 2.3.1.2. The dilution of the variant at 4% frequency was used as the normal counterpart, to simulate the possible presence of tumour contaminant in the normal counterpart.

Our procedure correctly detected the mutated base down to 20% frequency. Therefore, our pipeline can correctly detect homozygous and heterozygous somatic variants within samples with ≥20% and ≥40% tumour content, respectively. Notably, all samples used for mutational screening had >50% tumour content (Table 9).

DNA sequencing unit in Cogentech performed the DNA sequencing and Fabio Iannelli from Francesca Ciccarelli's group analysed the data.

Table 9: Variant frequency at different dilutions

Dilution	*Variant frequency (%)*	*Simulated contamination (homozygous variants)*	*Simulated contamination (heterozygous variants)*	*Total coverage*	*Variant coverage*	*Base Calling*
1	41.4	58.60%	17.20%	64	18	C
2	36.6	63.40%	26.80%	35	18	C
3	32	68.00%	36.00%	53	20	C
4	27.2	72.80%	45.60%	63	17	C
5	23.1	76.90%	53.80%	62	16	C
6	19.1	80.90%	61.80%	65	19	C
7	15	85.00%	70.00%	40	6	T
8 (Normal)	3.6	NA	NA	40	2	NA

For each set of dilution reported is the corresponding variant frequency, simulated contamination for homozygous and heterozygous, coverage after sequencing and the base identified at the position

2.2.4 SNP array for detecting copy number variations in human samples

We extracted the genomic DNA from the seven human tumour and their matched normal samples and processed according to Infinium® HD assay ultra manual (Table 5). We restored the DNA from FFPE samples before SNP array processing according to Infinium HD FFPE restore protocol. We assayed all samples using Illumina HumanOmniExpress-12 v1.0 and scanned the image data using a BeadArray reader. We extracted the intensity and genotype data for copy number variation analysis after normalizing raw fluorescent signals using Illumina Genome Studio v2011.1.

Agnese Collino from Gioacchino Natoli's group extracted the DNA and Microarray unit in Fondazione Filarete restored the FFPE DNA and processed the SNP array.

2.2.5 TaqMan copy number assay for copy number variation validation

We assessed copy number variation in *Mdr2*-KO HCCs by quantitative RT-PCR, using TaqMan Copy Number Assay, on a 7900HT Fast Real-Time PCR System (Applied Biosystems) with Sequence Detection Systems Software 2.2.2. We used TaqMan probes designed by the manufacturer for the experiment (Table 10) and *Tert* (Applied Biosystems, part number 4458373) as the reference. We plated all samples in quadruplicates with approximately 20 ng of DNA for each reaction. We analysed all TaqMan copy number assay experiments using CopyCaller v2.0 (Applied Biosystems). For each nodule, we used its matched normal tissue as the reference and considered a gene as amplified or deleted if the confidence scores associated with the copy number passed quality metrics as provided in the CopyCaller manual (http://www6.appliedbiosystems.com/support/software/copycaller/).

Real Time PCR Unit in Cogentech performed the quantitative RT-PCR.

Table 10: TaqMan probes used for validation of copy number amplification

Gene Name	*TaqMan Coordinate*	*Assay Id*	*Amplicon size (bp)*
Map2k7	Chr8:4238961	Mm00251746_cn	86
Irs2	Chr8:11005122	Mm00580766_cn	70
Mapk8	Chr14:34203892	Mm00418304_cn	95
ATF2	Chr2:73688980	Mm00046575_cn	101
MYC	Chr15:61821405	Mm00615587_cn	73
St18	Chr1:6817977	Mm00040629_cn	83
Hnf4a	Chr2:163377436	Mm00060776_cn	70
Gata1	ChrX:7539550	Mm00628417_cn	75
Gdf3	Chr6:122559836	Mm00561648_cn	70
Wnt10b	Chr15:98604678	Mm00612962_cn	81

For each gene reported is the corresponding coordinate in the mouse genome (mm9), Applied Biosystems TaqMan copy number assay identification number and the size of the amplicon in bp.

2.2.6 Fluorescence *in situ* hybridization

We validated the amplification of chromosome 19 in the human sample 23836 by two-colour fluorescence in situ hybridization (FISH) using a Vysis LSI 19q13 SpectrumOrange / 19p13 SpectrumGreen probe (Abbott), according to manufacturer's instructions. We deparaffinised 2 mm FFPE slides from tumour and background liver of patient 23286 in xylene, washed in 100% ethanol, incubated in 1x SSC (0.3M sodium chloride, 0.03M sodium citrate) pH 6.0 at 80°C for 20 min for demasking, and digested with pepsin (0.5 mg/ml in 0.2N HCl, pH 1.0; Protease and Protease Buffer II, Abbott) for 17 min at 37°C. We then washed samples in 2x SSC, dehydrated in 70, 95 and 100% ethanol, and directly applied air dried 10ul of probe onto each slide and topped with a coverglass that was then sealed with rubber cement. We placed the slides in a HYBrite (Abbott), and the probe was left to denature for 1 min at 85°C, followed by an overnight hybridization at 37°C. We then removed the coverglasses and washed the slides twice in 2x SSC with 0.1% NP-40 at RT, once in 0.4x SSC with 0.3% NP-40 at 73°C, and once again in 2x SSC with 0.1% NP-40 at RT. After counterstaining with DAPI (Sigma), we scored the FISH signals with an Olympus BX61 upright microscope, using a 100x objective.

Agnese Collino from Gioacchino Natoli's group prepared the slides for FISH experiment and Luca Giorgetti performed the FISH analysis.

2.2.7 Pathway enrichment analysis

We performed pathway enrichment analysis using ConsensusPathDB (Kamburov, et al. 2013). We compared 935 human cancer genes amplified in at least 4 BSEP-HCCs

and 27 genes that were amplified in the majority of human and mouse HCCs to the pathway-base gene set composed of 10,529 pathway-associated genes. ConsensusPathDB calculates p-values using the hypergeometric test based on the number of pathway components present in both the amplified cancer gene set and the pathway-base gene set. It then corrects the resulting p-values for multiple testing using false discovery rates.

Fabio Iannelli from Francesca Ciccarelli's group performed the pathway analysis.

2.2.8 Expression quantification of *Map2k7* in mouse liver cancer

We extracted total RNA for q-RT-PCR experiments from *Mdr2*-KO tumours, *Mdr2*-KO inflamed livers, and age matched *Mdr2*-WT healthy livers in Trizol (Invitrogen) using the RNeasy Mini Kit (Qiagen) according to manufacturer's instructions and used 0.5 ug of total RNA for cDNA synthesis (using ImProm-II Reverse Transcriptase, Promega). We then performed RNA quantification using Nanodrop, and assessed their quality using Bioanalyzer (Agilent). We quantified the expression by qPCR on 1 µl of cDNA reverse-transcribed from 0.5 µg of total RNA. We used Applied Biosystems 7500 Real-time PCR system to extract cycle threshold (Ct) values from qPCR (SYBR Green, Applied Biosystems). We then normalized the target gene Ct values with the nucleolin Ct values of the same sample to get delta Ct values for each gene in each samples. Next, we calculated the expression value as 2 to the power of negative delta Ct. For each gene with replicates; we then calculated the average expression for each gene.

Real Time PCR Unit in Cogentech performed the quantitative RT-PCR.

2.2.9 Treatment of *Mdr2*-KO mouse with SP600125 JNK inhibitor

We randomly divided twenty-three *Mdr2*-KO mice into two groups at the age of 13 to 14 months, when nodules are already formed (Pikarsky, et al. 2004). One group of 12

mice was treated with SP600125 (anthra[1,9-cd]pyrazol-6(2H)-one) (Calbiochem), and the other group of 11 mice with vehicle). Vehicle for SP600125 was 40% polyethylene glycol (PEG, Sigma) in PBS diluted in DMSO. We administered treatments of 60 mg per dose, intraperitoneally 3 times a week for a total of 3 weeks and sacrificed the mice one week after the end of the treatment. We counted all grossly detectable nodules and measured them with a calliper. We then collected them for DNA extraction and histological analysis.

Agnese Collino and Paola Nicoli from Gioacchino Natoli's group administered the drug, sacrificed the mice and extracted the DNA.

2.3 Computational Procedure

We performed all computational analysis in Francesca Ciccarelli's group, Department of Experimental Oncology, European Institute of Oncology (IEO), Milan, Italy.

2.3.1 Alignment and variant calling from targeted re-sequencing data

We mapped paired-end sequencing reads from each tumour and reference to the human genome (GRCh37/hg19) using Novoalign (http://novocraft.com). We allowed at most three mismatches per read and removed duplicated reads using rmdup of SAMtools (Li, et al. 2009). We considered all reads that uniquely mapped within 75-100 bp of the targeted regions as on target and retained them for further analysis (Table 11). We then identified single nucleotide variants (SNVs) and indels using SAMtools (Li, et al. 2009) and VarScan 2 (Koboldt, et al. 2012) and retained those variants that were covered by at least 10 reads and with frequency ≥20%. We next identified somatic mutations and indels as mutations with coverage ≥5x, frequency <10% in the reference, and not present in dbSNP build 137 (MAF >1%). All 44 SNVs and 8 indels were retained after manual inspection and 14 non-silent SNVs underwent orthogonal validation.

Table 11: Sequencing and alignment throughput of BSEP-HCCs

ID	*Sequenced Gbps*	*Aligned Gbps*	*Aligned w/o duplicates Gbps*	*On Target Gbps*	*Mean coverage*	*Mean coverage of matched normal*
175	17.15	15.38	11.62	8.17	160	157
7860	30.30	25.42	3.93	2.57	50	33
23836	13.46	11.96	8.27	5.52	108	110
HB4R	16.82	12.27	5.76	2.67	52	150
1790	8.70	5.99	5.45	4.73	92	146
2896	9.70	6.46	5.93	5.16	101	239

For each lesion of the human tumours reported are the sequenced bases, bases aligning to the mouse genome (mm9) before and after removing duplicates, bases aligning to the targeted regions of the genome, mean coverage in the tumour and the matched normal.

We mapped single- and paired-end reads to the mouse genome (NCBI37/mm9) using Novoalign (http://novocraft.com) for the targeted sequencing and using Burrows-Wheeler Aligner (BWA) (Li and Durbin 2009) for whole genome sequencing. We allowed a maximum of three mismatches per read and removed the duplicated reads generated during the PCR amplification using rmdup of SAMtools (Li, et al. 2009). In the targeted sequencing analyses, we considered all reads that uniquely mapped within 75-100 bp of the targeted regions as on target and retained these reads for further analysis (~80% of all aligned reads) (Table 12). For the whole genome sequencing, around 70% of raw reads aligned to the reference genome (Table 12). We identified mutations and indels using SAMtools (Li, et al. 2009) and VarScan 2 (Koboldt, et al. 2012) and retained variants covered by at least 10 reads with frequency ≥20%. We then identified somatic mutations and indels from the entire pool if they had frequency <5% in the reference sample and coverage of the mutation ≥10x. We used the same normal tissue as unique control because, in principle, mice from an inbred strain are genetically identical. We further manually inspected all the retained variants.

Fabio Iannelli from Francesca Ciccarelli's group performed the alignment and variant calling and mutational analysis.

Table 12: Sequencing and alignment throughput of Mdr2-KO HCCs

ID	*Target regions*	*Sequenced Gbps*	*Aligned Gbps*	*Aligned w/o duplicates Gbps*	*On Target Gbps*	*Mean coverage*	*Mean coverage of matched normal*
51509/1	Whole exome	16.44	10.50	10.43	6.34	126	132
60400/2		17.74	11.39	11.39	7.13	142	132
218/1		15.40	9.92	11.20	6.94	138	132
52686/1		21.05	13.47	12.54	7.78	154	132
58853/3		17.65	11.24	12.02	7.34	146	132
60400/1		17.72	11.54	10.06	6.26	124	132
58163/3		19.16	12.22	12.74	7.82	155	132
58163/4		15.52	9.99	9.61	5.31	105	132
215/1		18.89	12.34	12.57	7.99	159	132
54913/10	866 genes	2.88	2.20	2.15	1.28	481	189
54913/8		3.13	2.06	2.01	1.20	450	189
55481/10		3.14	2.25	2.20	1.27	476	189
55484/4		2.51	1.46	1.43	0.84	317	189
218/3	Whole genome	37.59	30.73	28.16	NA	11	13
60400/1		48.03	35.30	32.78	NA	13	12

For each lesion of the mouse tumours reported are the screening, sequenced bases, bases aligning to the mouse genome (mm9) before and after removing duplicates, bases aligning to the targeted regions of the genome, mean coverage in the tumour and the matched normal

2.3.2 CNV analysis from SNP arrays

We performed copy number variation (CNV) analysis using ASCAT (version 2.1), which takes into consideration ploidy and non-aberrant cell admixture present in each tumour sample (Van Loo, et al. 2010). For each genomic segment of the tumour, the software provides an aberration reliability score, which measures how well the predicted integer copy number matches the real data as compared to the null hypothesis of no aberration.

We ran ASCAT with default parameters of segment lengths for ASPCF Segmentation and homozygous probes (probes with 0.3>BAF>0.7 in matched reference) for all samples except two cases (1790 and 7860). For sample 1790, we masked the SNPs with BAF value difference <1.0 between tumour and reference to further reduce the noise and for sample 7860, we changed the segment length for ASPCF Segmentation from 25 to 100 to avoid over-segmentation due to scattered values of log ratio and BAF. We analysed all tumour samples by comparing them with their matched reference, except for HB4R. We analysed HB4R without the matched reference because the SNP array for the background liver failed quality control. In this case, we ran ASCAT using the Illumina700k platform provided in ASCAT for somatic CNV detection in absence of matched reference.

For six tumours with whole exome sequencing data, we integrated the frequency distributions of the germline heterozygous SNPs with the SNP array results to identify high confidence aberrant regions. In a diploid genome, heterozygous SNPs follow a unimodal distribution centred around 50% frequency because both alleles are present at equal frequency. In case of allelic imbalance leading to copy number variation, frequency distribution of heterozygous SNPs deviates from unimodality and their frequency will be different from 50%, because of the unbalanced ratio between mutated and wild type allele (Baca, et al. 2013). Hence, the distribution of heterozygous SNP frequencies can be used to identify genomic regions undergoing copy number variations. We identified high confidence aberrant regions as genomic segments with copy number different from 2 or with allelic imbalance leading to copy neutral loss of heterozygosity (CN-LOH) and with either an aberration reliability score higher than the average reliability score for that sample or present in regions with non-unimodal SNP frequency distribution. In the case of sample UKT, which did not undergo exome sequencing, we defined high confidence aberrant regions only on the basis of their aberration reliability score.

To identify amplified, deleted, and CN-LOH genes, we intersected the genomic coordinates of the aberrant regions in each sample with those of 20,965 human genes of

the SureSelect XT Human All Exon V4 kit (Agilent) and considered genes as modified if ≥ 80% of their length was contained in an aberrant region.

We further assessed the copy numbers of aberrant regions (amplifications, deletions, and CN-LOH) using ASCAT and merged adjacent regions with the same copy number. We defined single amplification, deletion, or CN-LOH events spanning ≤ 50% of the chromosome arm length as focal as opposite to arm-level alterations that spanned ≥98% of the chromosome arm length (Beroukhim, et al. 2010). We considered a minimum of ten consecutive copy number oscillations between two copy number states within the same chromosome as a possible indication of chromothripsis, according to its operational definition (Korbel and Campbell 2013).

I had performed the CNV analysis from SNP arrays.

2.3.3 CNV and structural rearrangement analysis from whole genome sequencing data

We performed copy number analysis on whole genome sequencing using CNVnator v.0.2.5 (Abyzov, et al. 2011) with sequence bins of 300bp. We then removed CNV regions with low confidence (p-value <0.05), composed of ≥65% repeats and/or gap, or covered with ≤30% of high quality mapping reads. Furthermore, we retained CNVs only if their length spanning at least 3 consecutive bins was >1000bp. We next identified somatic CNVs if there was ≤ 5% overlap in length between tumour copy number regions and matched normal. To identify the CNV boundary, we merged the adjacent regions with the same copy number state. We estimated the copy number fold change of each region of the tumour genome by dividing the tumour coverage by the coverage of the normal counterpart, after coverage normalization. We normalized sequence coverage (1) to reduce the coverage variations within the sample using the trimmed mean method (Dillies, et al. 2013) and (2) to scale the coverage of the tumour to its normal counterpart (Quackenbush

2002). To estimate copy number state in the tumour, we multiplied the fold change by 2, which is the expected copy number of a diploid genome. We considered a minimum of ten consecutive copy number oscillations between two copy number states within the same chromosome as a possible indication of chromothripsis, according to its operational definition (Korbel and Campbell 2013).

We inferred structural rearrangements using PEMer (Korbel, et al. 2009) with slight modifications to adapt the method to Illumina sequencing. We first calculated the paired-end insert size distribution to determine the expected insert size range and selected all mapped paired-end pairs with either an insert size greater/less than the expected insert size or with an unexpected orientation. Of these, we selected only discordant read pairs overlapping with or next to mapped CNV regions in the tumours for manual inspection. We identified rearrangements between chromosomes 8 and 14 (ID: 218/3) and 8 and 19 (ID: 60400/1), which we also validated by PCR amplifications and Sanger sequencing.

Fabio Iannelli from Francesca Ciccarelli's group analysed the WGS data for structural rearrangements and I had performed the CNV analysis.

2.3.4 GeneCNV

We developed a novel method, GeneCNV that identifies genes undergoing genomic alterations from whole exome and targeted sequencing data. We applied GeneCNV to identify altered genes in *Mdr2*-KO HCCs that underwent whole exome re-sequencing or targeted re-sequencing. In addition, we also assessed the performance of GeneCNV and compared its performance with three other state-of-the-art exome-based methods: ExomeCNV (Sathirapongsasuti, et al. 2011), EXCAVATOR (Magi, et al. 2013) and VarScan 2 (Koboldt, et al. 2012).

I had designed and developed the method. Gennaro Gambardella and Matteo Cereda contributed in the development of the software and tested the software. Fabio

Iannelli supervised the method and tested the software. I had performed the CNV analysis from targeted re-sequencing data using the three exome-based methods (ExomeCNV, EXCAVATOR and VarScan 2) and SNP arrays used as the gold standard for the assessment of the exome-based methods.

2.3.4.1 Workflow of the method

GeneCNV uses as an input exome sequencing data from two samples: the test and the reference sample. Although, in principle, these samples may be of any kind, for clarity, we refer to them as tumour and normal exomes from here on.

To detect altered genes from targeted and whole exome sequencing data, GeneCNV first calculates the coverage of targeted exons using BEDTools CoverageBed (Quinlan and Hall 2010) as:

$$ExonCoverage_e = \sum_{d=0}^{\max(depth)} d \times b_d$$

where d is the depth of coverage and b_d is the number of bases at depth of coverage d in exon e.

It next maps the exons to the corresponding gene using the targeted gene annotation file (i.e. the Agilent SureSelect bed files equivalent) and measures the gene coverage as the cumulative coverage of all exons divided by the length of the gene:

$$GeneCoverage_g = \frac{\sum_{e=1}^{no.\ of\ exons} ExonCoverage_e}{\sum_{e=1}^{no.\ of\ exons} ExonLength_e}$$

where e is the exon and g is the gene.

To minimize the variability of capture and sequencing efficiency within the exome, GeneCNV then applies median scaling normalization (Ioannidis, et al. 2009; Dillies, et al. 2013):

$$GeneCoverage_g' = \frac{\sqrt{GeneCoverage_g}}{median(\sqrt{GeneCoverage_g})_s}$$

where g is the gene and s is the exome sample.

Furthermore, it applies quantile normalization (Bolstad, et al. 2003) to correct for gene coverage variations between tumour and matched normal exomes. Genes in the tumour and in the matched normal samples are ranked according to their normalized gene coverage values (GeneCoverage'). The coverage of genes occupying equivalent positions in the two ranked lists is then reassigned as the average gene coverage between the two values (GeneCoverage'').

After normalization, GeneCNV calculates the log2ratio between the gene coverage in the tumour and in the matched normal exome ($L2R_{GC}$) for each gene:

$$L2R_{GC} = Log_2\left(\frac{GeneCoverage''_{g,\,tumor}}{GeneCoverage''_{g,\,normal}}\right)$$

In order to identify sample-specific thresholds of $L2R_{GC}$ for calling amplified and deleted genes, GeneCNV relies on the deviation from the expected 50% frequency of heterozygous SNPs in cases of allelic imbalance due to CNVs. GeneCNV first identifies heterozygous SNPs as germline mutations within 40-60% frequency interval in the normal sample. It then divides the tumour exome into non-overlapping regions with each region containing 100 such heterozygous SNPs and considers regions that host ≥80% of heterozygous SNPs within 40-60% frequency as allelic balanced regions. Some of these regions could however be polyploidy due to amplification of both the alleles. Since allelic balanced regions with polyploidy have high $L2R_{GC}$ values, GeneCNV considers regions with $L2R_{GC}$ values in the bottom 10% of the distribution of $L2R_{GC}$ values of all allelic balanced regions as diploid. It then finally calculates the thresholds for calling amplified and deleted genes from the distribution of $L2R_{GC}$ in the diploid regions as:

$$L2R_{GCA} = \overline{(L2R_{GC})}_{10\%} + 1SD_{(L2R_{GC})_{10\%}}$$

$$L2R_{GCD} = \overline{(L2R_{GC})}_{10\%} - 1SD_{(L2R_{GC})_{10\%}}$$

where SD represents standard deviation of the $L2R_{GC}$ values in the diploid regions.

We empirically estimated the optimal numbers of SNPs to divide the exome (100 SNPs), the $L2R_{GC}$ value to define diploid regions (10% of the distribution) and the number of standard deviations (SD). Though we have set theses as the default parameter values for running GeneCNV, the user can change these parameters values.

GeneCNV next uses SNP frequency and $L2R_{GC}$ of genes to detect copy number change. In order to assess the allelic copy number change, it uses a minimum of five SNPs present in the gene region. For genes that do not contain SNPs, GeneCNV adds 500bp to the flanking regions iteratively till the new gene region contains five SNPs. Since the new gene region may include other genes, GeneCNV re-calculates the $L2R_{GC}$ value for the gene as the mean of $L2R_{GC}$ values of all genes contained within the new gene region.

GeneCNV considers a gene to have allelic imbalance, if a minimum of two third of the SNPs do not have frequency between 40%-60%. It then identifies genes as amplified, deleted or LOH:

$$f(Gene) = \begin{cases} Amplified & if \;\; L2R_{GC} \geq L2R_{GCA}, \\ Deleted & L2R_{GC} \leq L2R_{GCD} \;\&\; allelic\ imbalance \\ LOH & L2R_{GCD} < L2R_{GC} < L2R_{GCA} \;\&\; allelic\ imbalance, \\ Wild\ type & otherwise \end{cases}$$

2.3.4.2 Dataset for exome-based method evaluation

To evaluate the performance of the exome-based methods, we used the six BSEP-HCCs for which the whole exome sequencing data and SNP array results were present for both tumour and matched normal samples. Furthermore, to increase the number of samples, we also included twenty-two samples from myelodysplasia (Yoshida, et al. 2011) for which whole exome sequencing data as well as SNP array data were publicly available for both tumour and their matched normal samples. We were thus able to do a comprehensive evaluation of the methods on a total of 28 tumour-normal pairs (Table 8).

Table 13: Exomes used for method comparisons

ID	*Read Length (bp)*	*Sequencing platform*	*Agilent SureSelect Human Kit*	*Targeted genes (n)*	*Alignment software*	*SNP array platform*	*SNP call rate in tumour*	*SNP call rate in matched normal*	*Tumour type*
CMML-01	76	Illumina GAIIx	All Exon 38Mb kit	19,104	BWA	Affymetrix 250K SNP arrays	99.13	98.86	Myelodysplasia (Yoshida, et al. 2011)
CMML-03	108		All Exon 50Mb kit	20,965			98.70	98.64	
CMML-04	108		All Exon 50Mb kit	20,965			98.00	98.31	
MDS-03	101 / 76		All Exon 38Mb kit	19,104			99.07	99.48	
MDS-04	101 / 76		All Exon 38Mb kit	19,104			98.72	99.54	
MDS-06	101		All Exon 38Mb kit	19,104			97.94	98.11	
MDS-07	101		All Exon 38Mb kit	19,104			96.90	98.81	
MDS-08	101 / 76		All Exon 38Mb kit	19,104			99.58	98.85	
MDS-09	108		All Exon 50Mb kit	20,965			98.72	99.53	
MDS-10	108		All Exon 50Mb kit	20,965			99.34	98.41	
MDS-13	108 / 76		All Exon 38Mb kit	19,104			97.57	98.02	
MDS-14	108 / 76		All Exon 38Mb kit	19,104			97.76	99.36	
MDS-15	108		All Exon 50Mb kit	20,965			99.25	95.80	
MDS-18	76		All Exon 38Mb kit	19,104			99.55	99.20	
MDS-20	108		All Exon 38Mb kit	19,104			99.54	99.34	

ID	Read Length (bp)	Sequencing platform	Agilent SureSelect Human Kit	Targeted genes (n)	Alignment software	SNP array platform	SNP call rate in tumour	SNP call rate in matched normal	Tumour type
tAML-01	101	Illumina GAIIx	All Exon 38Mb kit	19,104	BWA	Affymetrix 250K SNP arrays	98.18	96.22	Myelodysplasia (Yoshida, et al. 2011)
tAML-02	108 / 101		All Exon 38Mb kit	19,104			98.91	97.00	
tAML-03	108 / 76		All Exon 38Mb kit	19,104			97.29	98.24	
tAML-04	108 / 76		All Exon 38Mb kit	19,104			97.93	97.72	
tAML-05	108		All Exon 38Mb kit	19,104			99.18	99.52	
tAML-06	108 / 101		All Exon 50Mb kit	20,965			98.76	98.58	
tAML-07	108		All Exon 50Mb kit	20,965			97.65	96.79	
175	101	Illumina HiSeq2000	All Exon 50Mb kit	20,965	Novoalign	Illumina HumanOmniExpress-12 v1.0	0.93	1.00	BSEP-HCCs
1790	101		All Exon 50Mb kit	20,965			1.00	1.00	
2896	101		All Exon 50Mb kit	20,965			1.00	0.99	
7860	76		All Exon 50Mb kit	20,965			0.73	0.86	
23836	76		All Exon 50Mb kit	20,965			0.95	0.98	
HB4R	101		All Exon 50Mb kit	20,965			0.84	0.87	

For each sample reported are the ID as in the original study; the length of sequenced reads (all experiments were performed with a paired end setting); the sequencing platform, the exome capture kit; the number of targeted genes, the software used for read alignment; the SNP array platform; the SNP call rate in the tumour and in the matched normal; and the tumour type.

2.3.4.3 Identification of altered genes from SNP array

We downloaded the raw SNP array files for myelodysplasia from Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/, GSE31174) (Yoshida, et al. 2011) and analysed them using ASCAT (version 2.1) (Van Loo, et al. 2010) with default parameters (segment length = 25; homozygous probes with 0.3>B allele frequency (BAF) >0.7 in matched reference were masked). High confidence somatic CNVs and CN-LOHs were identified in each tumour sample as genomic segments with an aberration reliability score higher than the average reliability score for that sample. We then identified altered genes in each sample by intersecting the genomic coordinates of the aberrant regions in each sample with those of the human gene sets of the Agilent SureSelect Human All Exon 50Mb kit (20,965 genes) or 38Mb kit (19,104 genes) depending on the kit used to capture the exome in the corresponding sample (Table 13). We considered genes as modified if ≥ 80% of their length was contained in an aberrant region. For the six samples from BSEP-HCCs, we used the SNP array results after processing the data as previously described. In total, we identified altered genes from SNP arrays of 28 samples (Table 13) that were then used as gold standard for comparison of exome-based methods.

2.3.4.4 Identification of modified genes from exome sequencing data

We downloaded the whole exome sequencing data for myelodysplasia from Sequence Read Archive (SRA, http://www.ncbi.nlm.nih.gov/sra , DRA000433) (Yoshida, et al. 2011). We mapped the sequencing reads from each tumour and matched normal to the human genome (GRCh37/hg19) using BWA (Li and Durbin 2009) allowing at most three mismatches per read and removed the duplicated reads using rmdup of SAMtools (Li, et al. 2009). We considered all reads uniquely mapping within 75-100 bp of the targeted regions as on target and retained them for further analysis. For the six BSEP-HCCs, we used the aligned files obtained after processing the reads as described.

We identified CNVs from exome sequencing data for each of the twenty-eight tumour exomes using GeneCNV, ExomeCNV (version 1.4), VarScan 2 (version 2.3.6) and EXCAVATOR (version 2.2). We ran all the methods using the default parameters (ExomeCNV: minimum sensitivity and specificity = 0.9999 and optimizing for specificity; VarScan 2: minimum coverage = 8, minimum size = 10 bases, log ratio threshold = 0.25 for both the lower and upper bounds; EXCAVATOR: mode = somatic). Since ExomeCNV, VarScan 2 and EXCAVATOR identify CNV regions, to detect deleted and amplified genes in each tumour exome, we performed similar analysis as described before for the SNP arrays. CN-LOH genes could only be identified using ExomeCNV and GeneCNV.

2.3.4.5 Comparison of GeneCNV with other methods

We considered the amplified, deleted and CN-LOH genes from the SNP arrays as the true alterations occurring in the tumour samples and used these results to compare the performance of the four exome-based methods. For each exome-based method, we defined true positives as the genes with the same alterations as detected in the SNP arrays and true negatives as unaltered genes in both. We then calculated sensitivity as the number of true positives over the total number of altered genes in the SNP arrays and specificity as the number of true negatives over the total number of unaltered genes in the SNP arrays. We also measured accuracy as the number of correct calls (sum of true positives and true negatives) over the total number of targeted genes. We next measured the concordance between the results from exome-based methods and SNP arrays using the Jaccard index as the number of true positives over the sum of altered genes detected by the exome-based method and in SNP arrays.

We also assessed the performance of the exome-based methods in detecting clonal events, where we identified clonal events using the BAF and the log R ratio (LRR) profile

from the SNP array. In particular, we considered altered regions with all SNPs in the tumour with the BAF value >0.6 or <0.4 or with the LRR value of >0.25 or <-0.25 as clonal events.

Results

The aim of my project is to understand the alterations that occur during liver cancer progression in the absence of external mutagens. HCC is a multistep process and arises in response to various external factors such as virus infection, aflatoxin and alcohol, or as a consequence of metabolic diseases including obesity and diabetes (Block, et al. 2003; El-Serag and Rudolph 2007). HCC is heterogeneous in terms of the genetic make up (Unsal, et al. 1994; Dragani 2010) and the acquired mutation signature depends on the initiating agent (Zhang 2012). Irrespective of the causative agents, HCC develops in the background of cirrhosis, fibrosis and chronic inflammation as a result of liver injury and regeneration (Levrero 2006). The contribution of chronic inflammation and cirrhosis to liver injury resulting in cancer is not well understood. Hence we investigated to which extent liver injury due to inflammation, chemical damage, and fibrosis aids in the acquisition of cancer genomic instability.

For the purpose of our study, we used human BSEP-HCCs, which have inherited inactivating mutations in the biliary transporter gene, *ABCB11* (BSEP). BSEP-HCCs are generally deficient in BSEP expression and develop in response to chronic liver injury due to the accumulation of bile salts (Knisely, et al. 2006). To further investigate the role of the acquired alterations in the progression of the human tumours, we used a genetic mouse model *Mdr2*-KO, which is etiologically similar to BSEP-HCCs. *Mdr2*-KO mice, like human BSEP-HCC patients have inactivating mutations in a biliary transporter gene, *ABCB4* and develop HCCs due to hepatocellular damage and inflammation (Smit, et al. 1993; Mauad, et al. 1994).

We profiled the genome of seven human and fourteen mouse liver tumours (Table 5 and Table 6). To study the mutations and CNVs profiles in both human and mouse tumours, we used known methods for identifying mutations and CNVs from SNP arrays and WGS data. We in addition developed a new method, GeneCNV to detect CNVs from targeted re-sequencing data. We applied GeneCNV on WES screenings in mouse tumours.

In the following paragraphs I first describe the rationale, validation and salient features of GeneCNV, followed by the analytical observations on the genomic alterations in human BSEP-HCCs and mouse *Mdr2*-KO HCCs.

3.1 GeneCNV: Rationale and performance evaluation

3.1.1 Rationale of GeneCNV

The detection of CNVs from NGS data is possible using read-depth or read-pair approach. Read-pair approach is based on mapping and orientation of reads at the breakpoints of the CNVs. This approach is not suitable for WES data because CNVs that have breakpoints in untargeted regions will be missed. Read-depth approach assumes that amplified regions have higher coverage and deleted regions have lower coverage in comparison to unaltered regions. This approach is widely used for detecting CNVs from WGS and WES data. The application of the read-depth approach to WES data is particularly challenging because of the sparse distribution of exons of variable length and sequence composition across the genome. This further affects the sequence coverage. Few methods such as ExomeCNV, VarScan 2 and EXCAVATOR have been developed in recent times for detecting CNVs from WES data. ExomeCNV compares the read depth of exons between tumour and normal samples and employs CBS, a segmentation method to identify breakpoints of copy number change in the tumour. However it does not take into consideration the difference in the total sequencing coverage between the normal and the tumour samples that may lead to ambiguous CNV calls. Similar to ExomeCNV, VarScan 2 exploits the coverage differences between tumour and normal samples and uses CBS to detect CNV breakpoints after normalizing for the coverage between the two samples. EXCAVATOR uses a new segmentation algorithm that exploits the distance between the adjacent exons, thus accounting for the sparseness of the exons in WES data. Though all the available methods use different normalization and CNV calling techniques, they

invariably use segmentation to identify regions of copy number changes. Segmentation has been widely used in identifying CNVs in WGS and microarrays, where information across the whole genome is available. However, segmentation had reduced efficacy in detecting CNVs from WES data because of the non-contiguous distribution of exons. Additionally, these methods do not take into consideration the effect of large-scale rearrangements, which is a common feature in cancers.

To address these issues, we developed a novel method that does not depend on segmentation and instead calls CNVs at the gene level. GeneCNV adopts a novel strategy to minimize the coverage variability because of exon length, composition, and discrete distribution along the genome that is different from those of other exome-based methods (Figure 19).

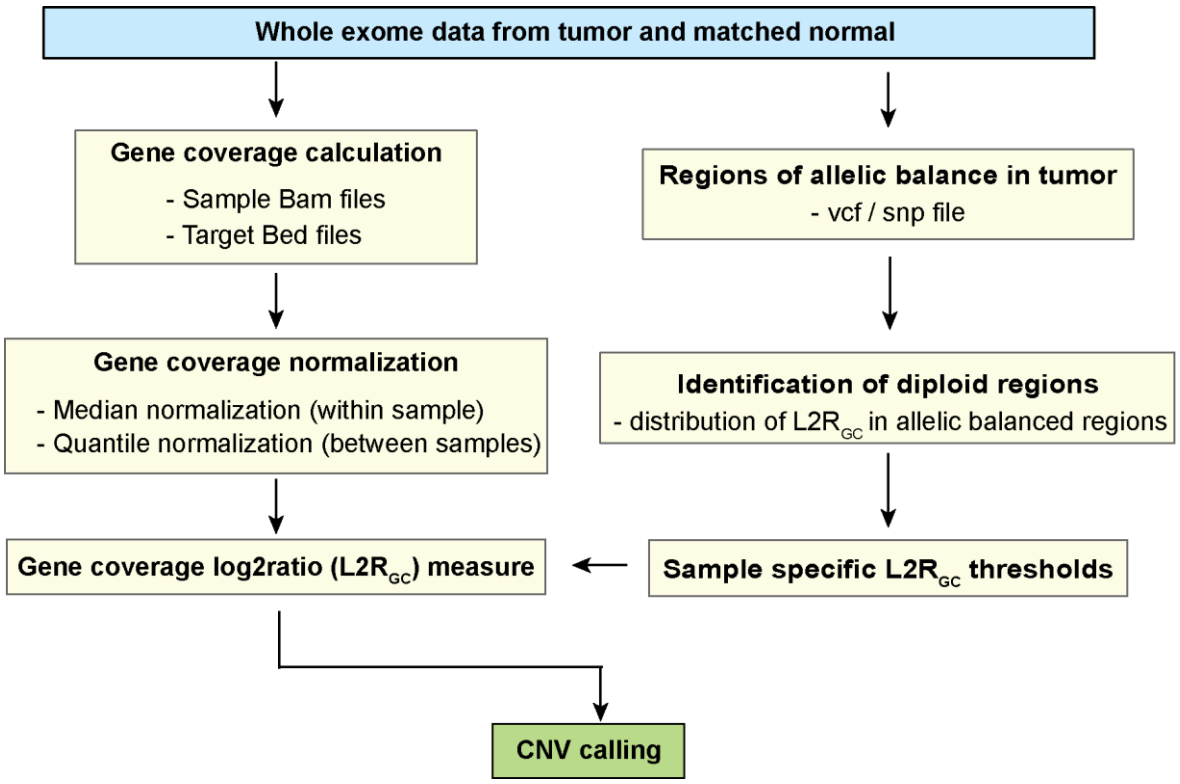


Figure 19: Schematic representation of GeneCNV

Shown is the schematic workflow of GeneCNV for detecting CNVs and CN-LOHs from whole exome sequencing data. It uses as inputs, aligned read files (BAM), genomic coordinates of regions targeted for re-sequencing (bed) and variant file (vcf or snp). It uses read aligned files of the

exomes to calculate the coverage of all targeted genes in test (*e.g.* tumour) and reference (*e.g.* matched normal) exomes independently. It normalizes the gene coverage within and between the samples using median normalization and quantile normalization, respectively. It then measures the $\log_2$ratio of gene coverage ($L2R_{GC}$) between tumour and matched normal exomes for each targeted gene and identifies the regions of allelic balance in the tumour. Within the allelic balanced regions, GeneCNV further identifies regions that maintain a diploid status in the tumour and defines $L2R_{GC}$ thresholds for amplifications and deletions. This allows the detection of CNVs, CN-LOH and two copy genes along the whole exome.

Exons of different length and sequence in the genome affect the sequencing efficiency, which leads to variability in exon coverage. In order to address this problem, GenCNV first merges all targeted exons of each gene to rebuild a contiguous region that spans the entire length of the gene. This is then used to calculate the average gene coverage, thus allowing uniformly captured and sequenced exons to compensate for the non-uniform coverage of other exons within the same gene (Figure 20).

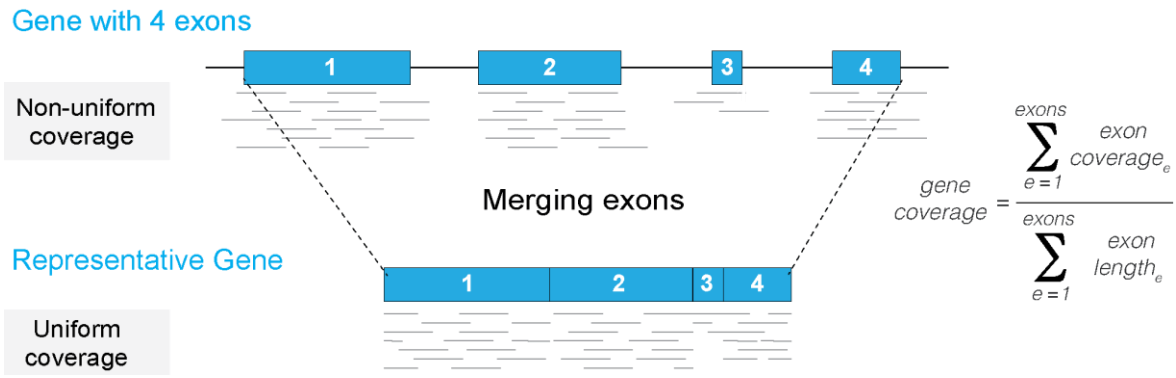


Figure 20: Gene coverage calculation

Shown is the step for calculating gene coverage. The coverage of each gene is calculated as the sum of coverage of the exons over the length of gene defined as the sum total of the length of exons constituting the gene. Merging the exons for a gene reduces the effect of non-uniform converge of exons (exon 3).

Further, coverage variation may still exist between genes within the same exome due to the sequence composition. To remove this bias, GeneCNV normalizes the gene coverage within the same exome using global median normalization (Figure 21). Next, it normalizes gene coverage across tumour and matched normal exomes using quantile normalization (Figure 21). Quantile normalization is an important step that accounts for the coverage variability between two samples due to different DNA quality, library preparation protocols, sequencing settings and performance.

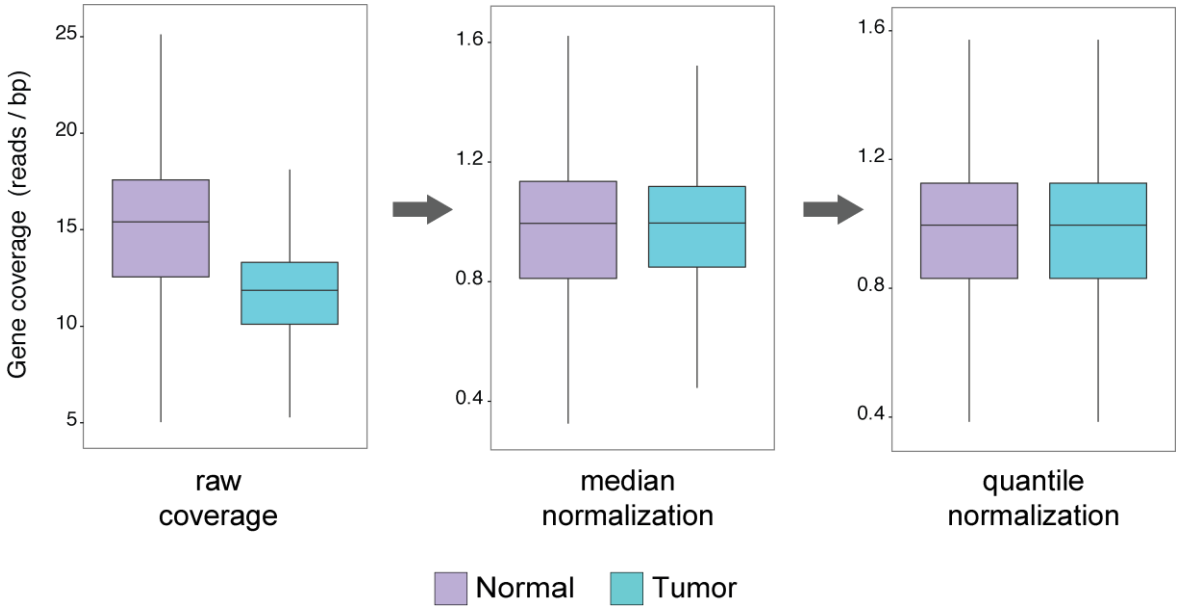


Figure 21: Distribution of gene coverage at each step of normalization

Shown are distribution of gene coverage at each the steps of normalization to remove variations due to technical issues related to WES data: length and sequence composition of the targeted exons, DNA quality, library preparation protocol, sequencing settings and performances. Median normalization is used to remove the coverage variations present within the exome and quantile normalization is used to remove the coverage variability between the samples.

To detect changes in copy numbers, GeneCNV exploits the difference in sequencing coverage between tumours and normal counterparts. It uses the normalized gene coverage in the tumour and normal exomes to calculate the gene coverage log_2ratio ($L2R_{GC}$) between tumour and matched normal for each targeted gene (Figure 22). In principle, $L2R_{GC}$ values around zero are indicative of genes with the same copy number in the tumour and in the normal counterpart, while $L2R_{GC}$ values higher than 0.6 (corresponding to ≥1.5 fold change) or lower than -1 (corresponding to ≤0.5 fold change) indicate tumour-specific gene amplifications and deletions, respectively (Figure 22A). However, such fixed thresholds can be used only in genomes where no widespread chromosomal rearrangements occur. Since most cancer genomes acquire significant chromosomal instability, they often show substantially modified $L2R_{GC}$ spectra (Figure 22B).

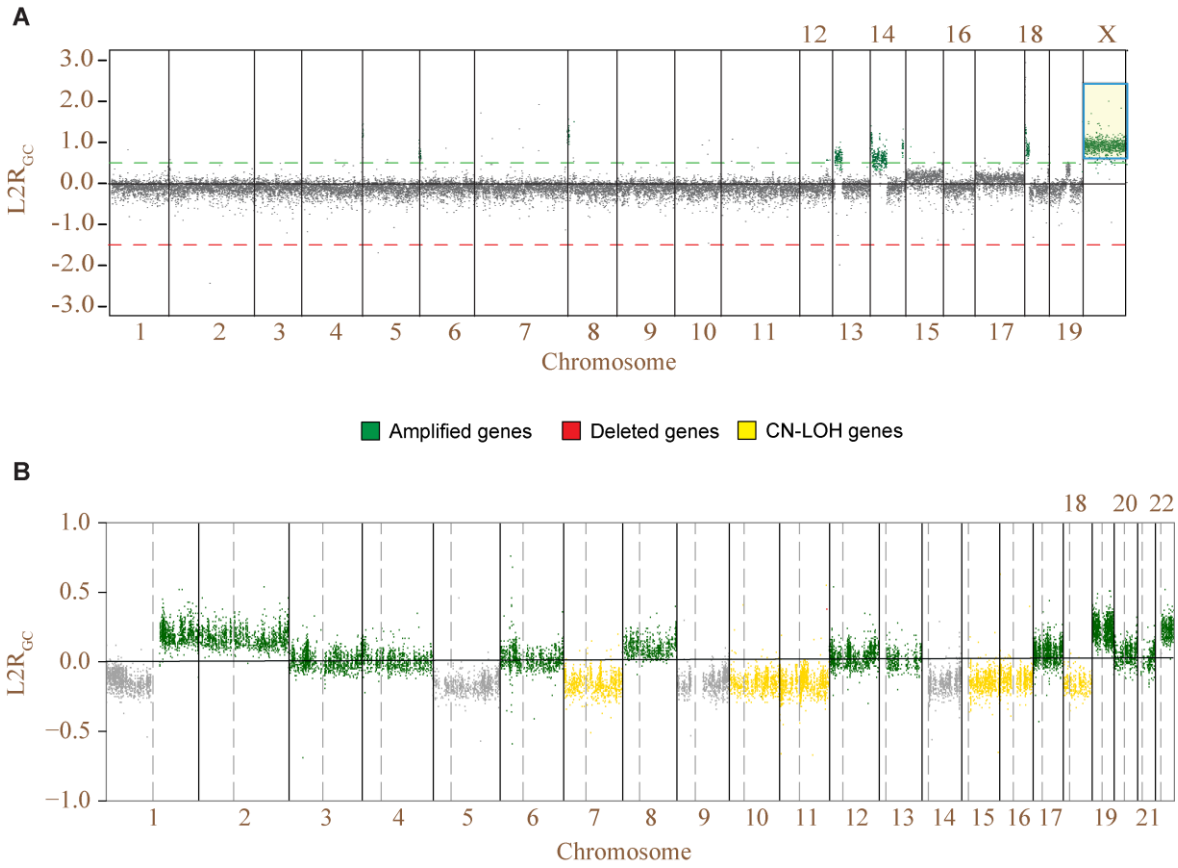


Figure 22: Gene coverage log_2ratio measure ($L2R_{GC}$) spectrum

Shown is the log_2ratio ($L2R_{GC}$) of coverage for each gene targeted in the whole exome sequencing data, calculated as the log ratio value between the normalized coverage of tumour over the normal.

A) The $L2R_{GC}$ is calculated between a normal sample from male and tumour sample from female of *Mdr2*-KO mice. Genes not undergoing CNVs have same coverage in both tumour and normal (grey) and hence have $L2R_{GC}$ values around zero (black lines). $L2R_{GC}$ of 0.6 corresponds to ≥ 1.5 fold change in copy number (green dashed line) and of -1.0 corresponds to ≤ 0.5 fold change in copy number (red dashed line). Additional copy of chromosome X in female compared to male corresponds to $L2R_{GC}$ >0.6. Genes are amplified (green) if $L2R_{GC}$ >0.5 and deleted (red) if $L2R_{GC}$ <-1.0 **B**) Substantially modified spectra of $L2R_{GC}$ in a human WES cancer sample, where $L2R_{GC}$ for unaltered genes is substantially shifted from expected zero value. Amplified (green), deleted (red), CN-LOH (yellow) and unaltered genes, two-copy (grey) are coloured according to results from SNP arrays.

For this reason, GeneCNV does not apply fixed $L2R_{GC}$ thresholds and instead derives them for each tumour sample by identifying the distribution of $L2R_{GC}$ in the diploid regions of the tumour genome. It first identifies regions of the tumour genome that maintain their allelic balance. In a normal diploid genome, allelic balance is maintained and the expected frequency of heterozygous germline mutations (SNPs) is around 50%. In the case of an allelic imbalance with different numbers of copies of the alleles, the frequency of the heterozygous SNPs deviates from the expected 50% (Figure 23). Hence, GeneCNV identifies regions of the tumour genome where the frequency of heterozygous SNPs is around 50%, thus indicating that their allelic balance is maintained.

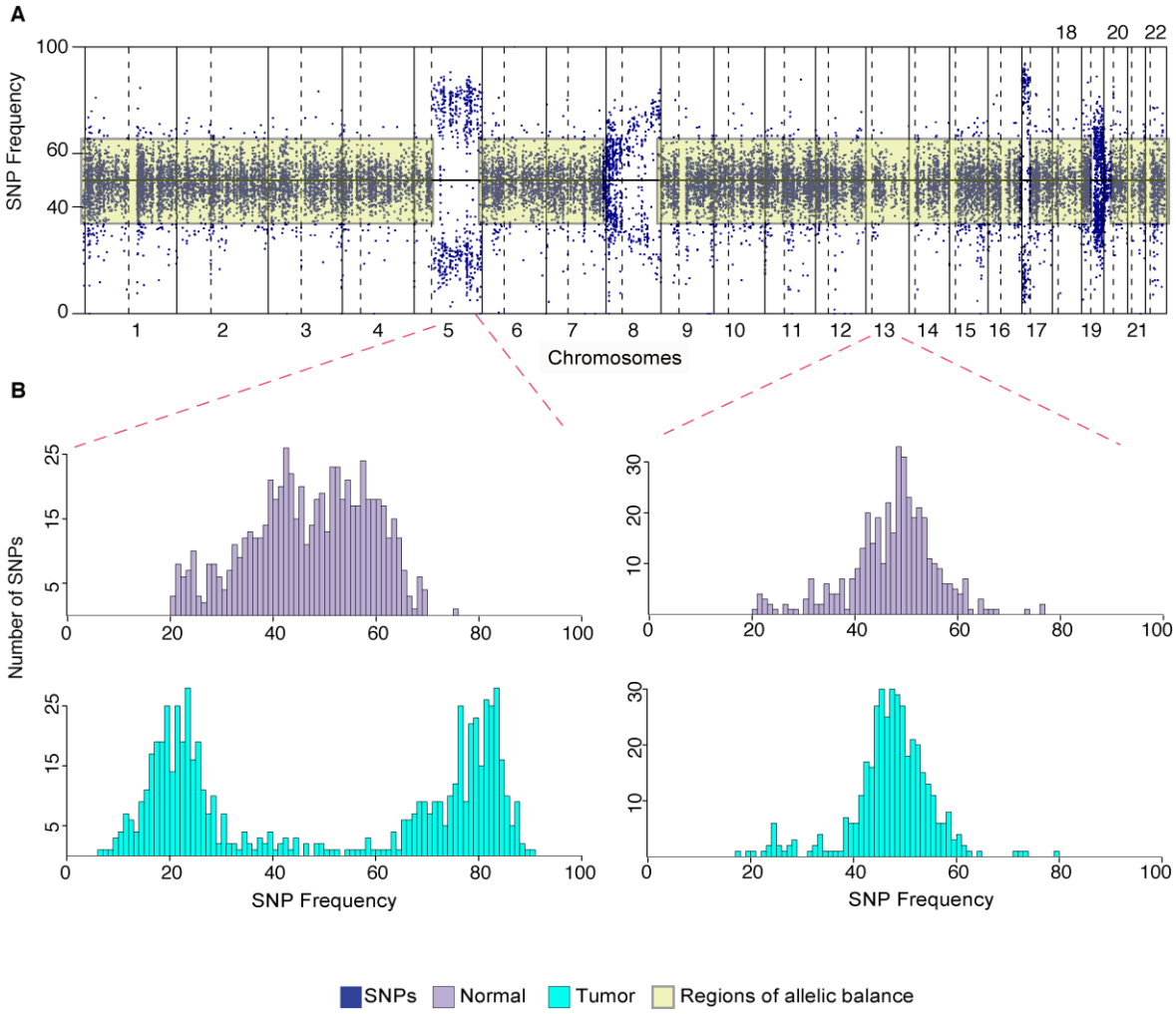


Figure 23: Identification of regions of allelic balance using SNP frequency

A) SNP frequency in human tumour sample. Regions of chromosome 5, 8, 18 and 19 undergo allelic imbalance whereas in other regions (yellow) allelic balance is maintained. B) Regions of allelic imbalance chromosome 5 and allelic balance of chromosome 13. Histogram shows the observed SNP frequency in normal and tumour conditions. As expected, for diploid region, the SNP frequency is around 50%, while it deviates from 50% in case of allelic imbalance.

GeneCNV then identifies diploid regions within these allelic balanced regions, as the ones with $L2R_{GC}$ values in the lower tail of the $L2R_{GC}$ distribution (Figure 24), because regions of allelic balance but with high $L2R_{GC}$ values may correspond to tumour-specific amplifications of both alleles. GeneCNV uses the distribution of $L2R_{GC}$ values in diploid

regions to identify sample-specific $L2R_{GC}$ thresholds for amplification ($L2R_{GCA}$) and deletion ($L2R_{GCD}$,) (Figure 24).

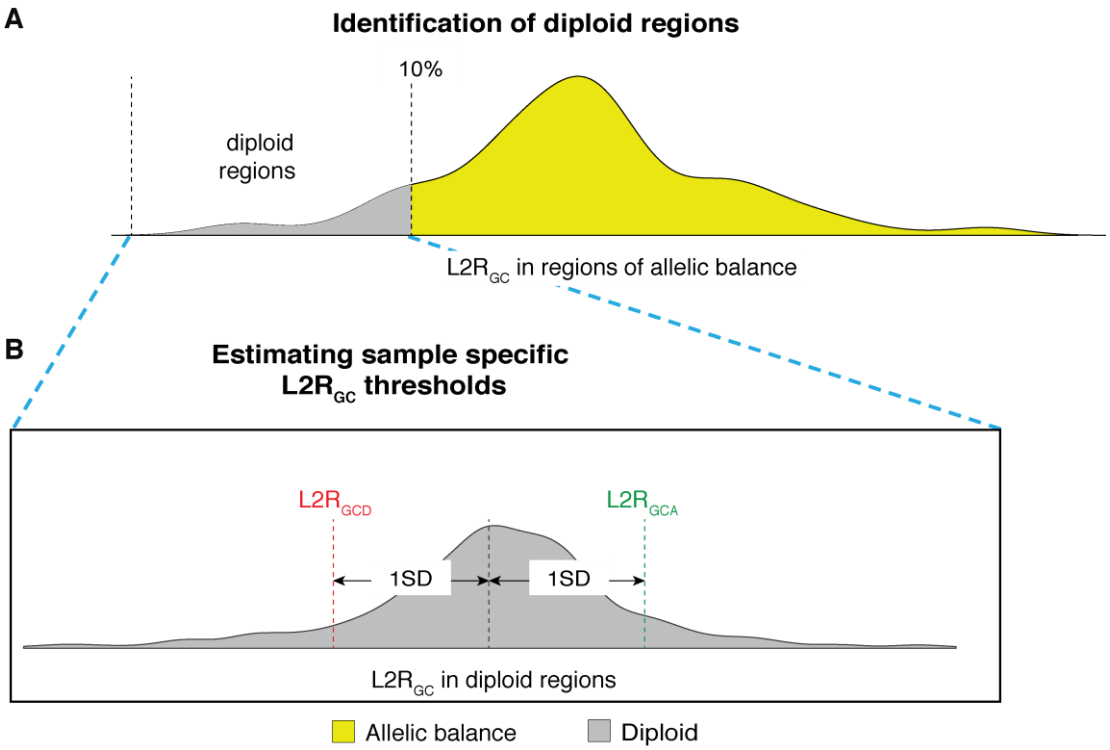


Figure 24: Identification of diploid region and estimating sample-specific thresholds

A) $L2R_{GC}$ distribution in allelic balanced regions. Lower 10% of the distribution corresponds to diploid region (grey) that have lower values of $L2R_{GC}$ when compared to regions with amplification of both alleles (yellow). B) Estimation of sample-specific thresholds for deletion ($L2R_{GCD}$) and amplifications ($L2R_{GCA}$), calculated as -1 SD and +1 SD from the mean of the $L2R_{GC}$ distribution in the diploid region respectively. SD=standard deviation

Genes in regions of allelic imbalance and with $L2R_{GC}$ values within the thresholds are considered as undergoing CN-LOH. All the other genes are regarded as maintaining a two-copy status (Figure 25).

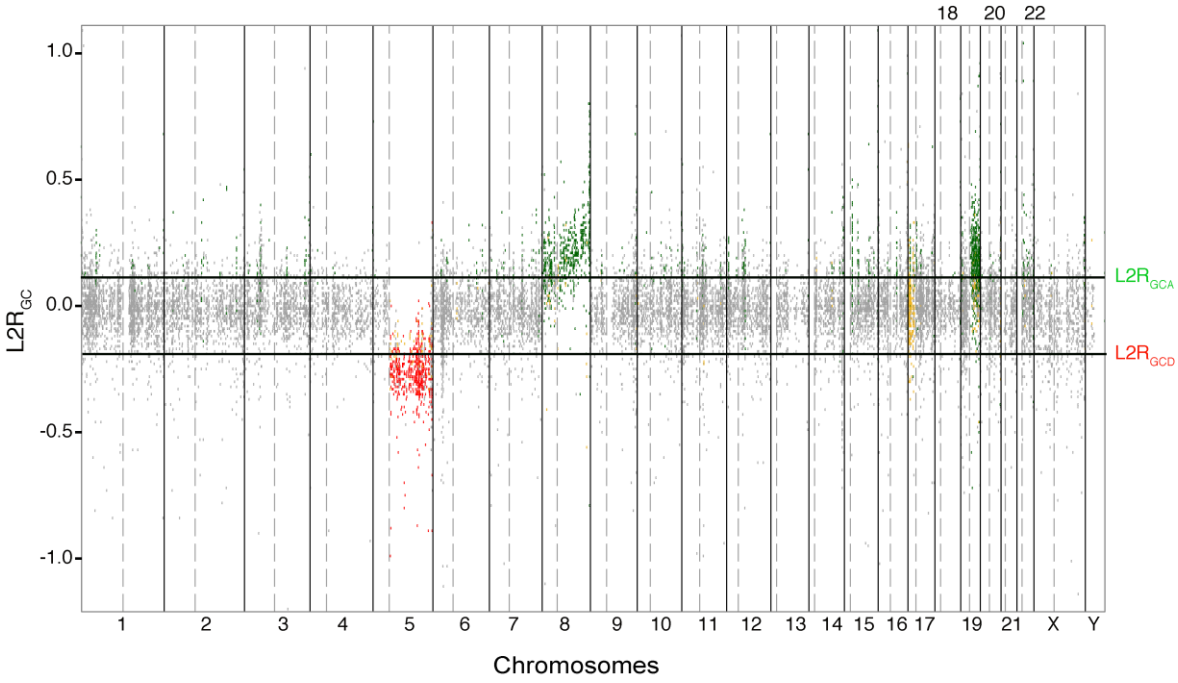


Figure 25: CNV calling

Shown is the result of GeneCNV in calling amplified (green), deleted (red) and CN-LOH (yellow) genes. Genes are identified as amplified if $L2R_{GC} \geq L2R_{GCA}$, deleted if $L2R_{GC} \leq L2R_{GCD}$ and CN-LOH if $L2R_{GCD} < L2R_{GC} < L2R_{GCA}$ and present in allelic imbalanced region.

3.1.2 Comparison of GeneCNV with other methods

We assessed the performance of GeneCNV as compared to three widely used exome-based methods, namely ExomeCNV, VarScan 2, and EXCAVATOR. As a test dataset, we used tumour and their matched normal WES data from 22 myelodysplasias (Yoshida, et al. 2011) and the six BSEP-HCCs (). Tumour-specific amplified, deleted and CN-LOH genes detected from the SNP arrays on the same samples were used as the gold standard for assessment. For each of the four exome-based methods, we measured sensitivity, specificity, accuracy, and the Jaccard index, which estimates the concordance with the SNP array results. Overall, GeneCNV showed the highest sensitivity (51%), accuracy (78%), and Jaccard index (36%) as compared to the other three methods (Figure 26). In addition, although it had the second best specificity (88%), it showed the best trade-off between specificity and sensitivity (Figure 27).

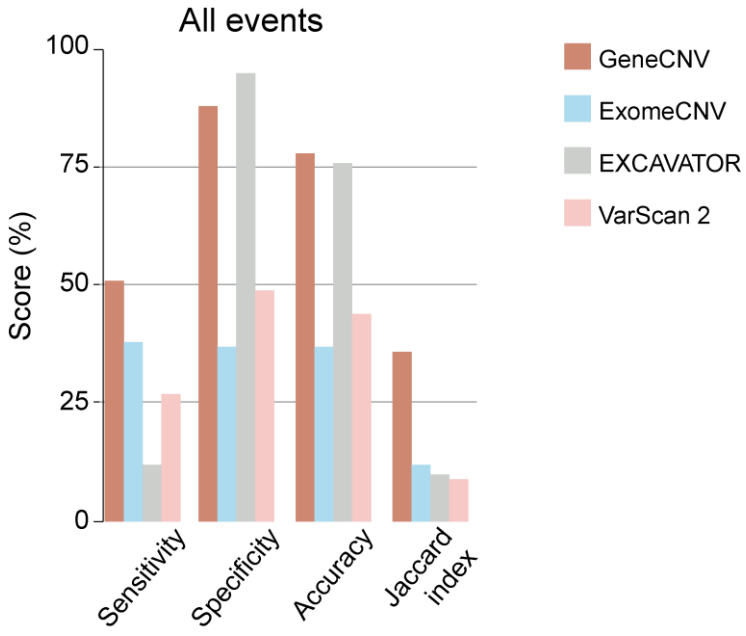


Figure 26: Comparison of GeneCNV with other exome-based methods for CNV detection in 28 tumour exomes

Reported are sensitivity, specificity, accuracy and Jaccard index of the four exome-based methods (GeneCNV, ExomeCNV, VarScan 2, and EXCAVATOR) in detecting all CNVs in the 28 samples as compared to SNP array

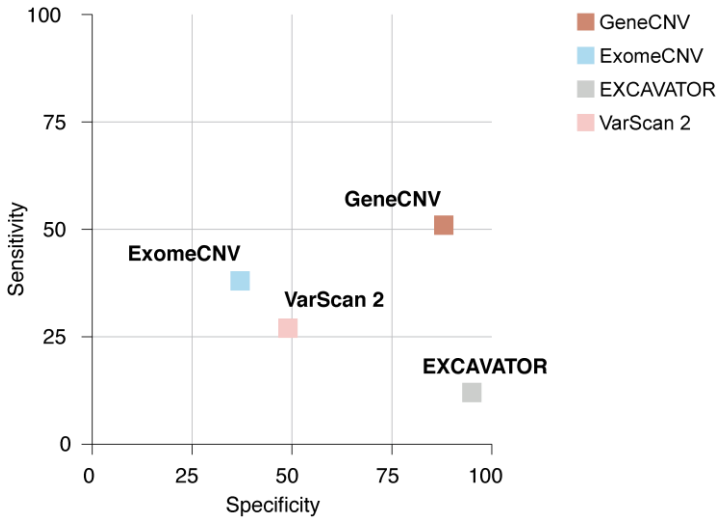


Figure 27: Trade-off between sensitivity and specificity for the four exome-based methods

Reported are the trade-off between sensitivity and specificity of each exome-based method in detecting CNVs in the 28 tumour exomes

In general, all four exome-based methods and in particular GeneCNV, were more sensitive in detecting amplified genes (Figure 28A) than deleted genes (Figure 28B). GeneCNV again showed the highest concordance with the SNP array results in detecting deletions (Figure 28B) whereas ExomeCNV and VarScan 2 showed high sensitivity but poor specificity, thus suggesting that the higher sensitivity of the other methods was due to an overall overestimation of deletions. Of the four exome-based method, only GeneCNV and ExomeCNV can detect CN-LOHs, and GeneCNV showed the best performance in all comparisons (Figure 28C).

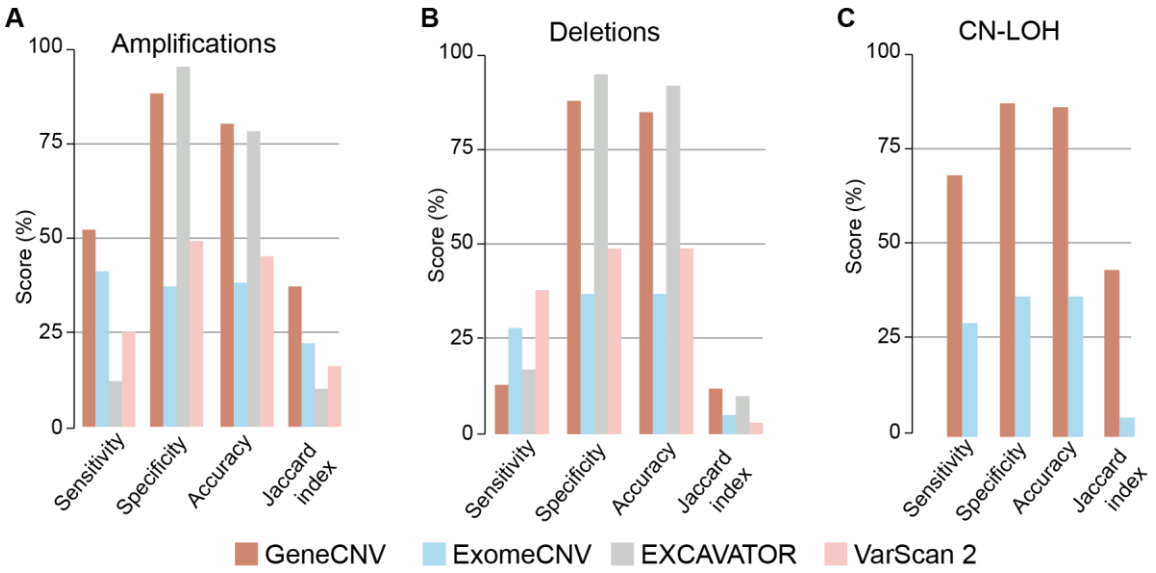


Figure 28: Performance assessment of the four exome-based methods in detecting amplifications, deletions and CN-LOH

Reported are sensitivity, specificity, accuracy and Jaccard index of the four exome-based methods (GeneCNV, ExomeCNV, VarScan 2, and EXCAVATOR) in detecting (A) amplified genes, (B) deleted genes and (C) CN-LOH.

Overall, exome-based methods had poor concordance with somatic copy number events called by SNP arrays (Figure 26). In order to understand the reasons for this, we assessed the performances in each tumour exome individually and noticed that exome-based methods consistently failed to detect any CNVs in some tumours (Figure 29).

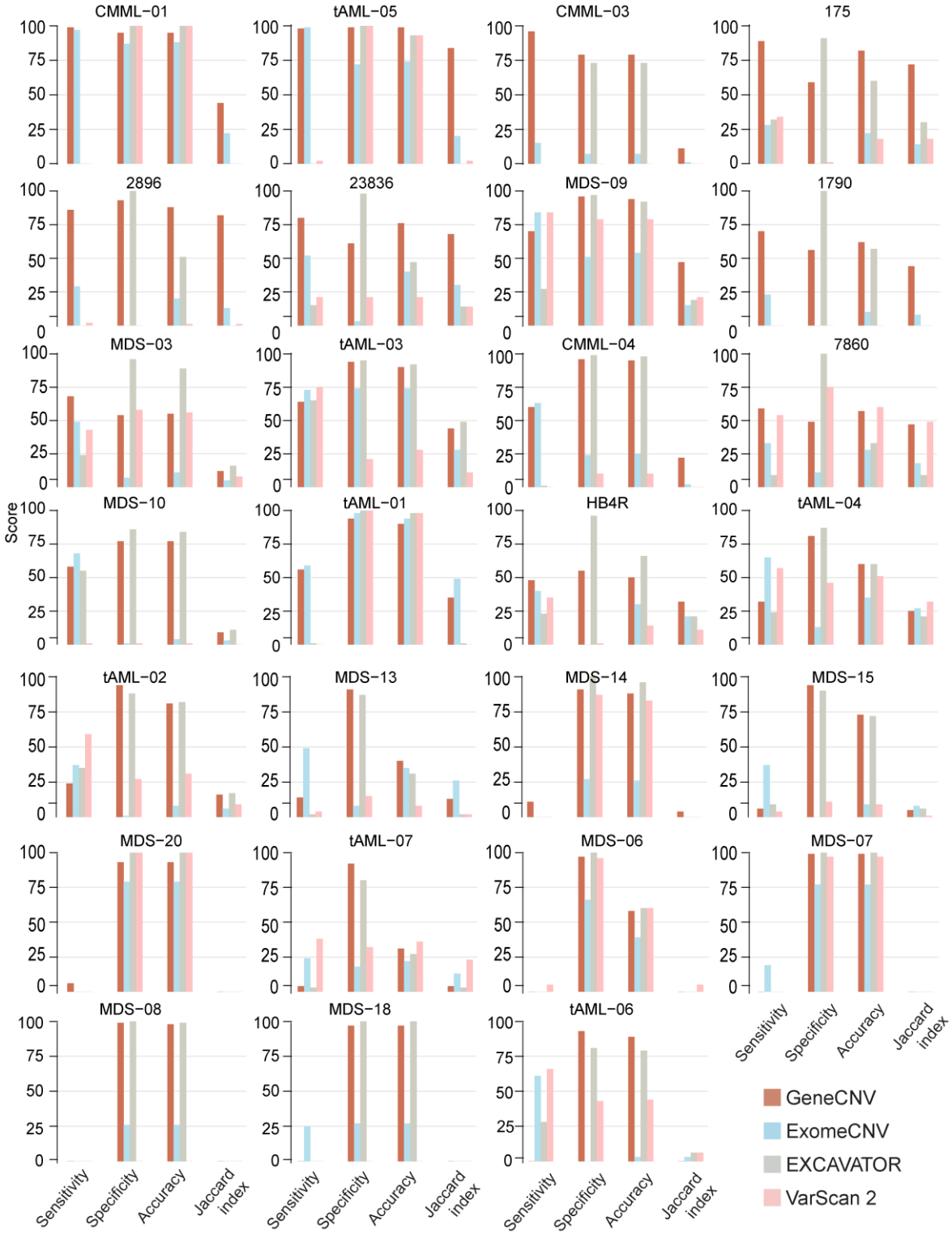


Figure 29: Performance assessment of the four exome-based methods in detecting altered genes in each 28 samples

Reported are sensitivity, specificity, accuracy and Jaccard index of the four exome-based methods (GeneCNV, ExomeCNV, VarScan 2, and EXCAVATOR) in detecting CNVs in each sample. Samples with at least one altered gene are shown.

There are multiple and probably concurrent reasons for the poor performance of exome-based methods. One possible reason could be the poor SNP call rate of some of the SNP arrays (e.g. MDS-15, tAML-07 and tAML-02, (Table 13)), thus suggesting the occurrence of possible false positive calls from the SNP arrays. Moreover, tumour somatic alterations are usually a mixture of clonal and subclonal events, depending on when they occur during cancer growth. Sensitivity of exome-based method is lower in detecting subclonal CNVs, because the differences in coverage between tumour and normal samples are not high enough to identify changes in copy number. This leads to false negative calls from exome-based methods. To understand whether this was the case, we defined clonal events based on the B allele frequency and log R ratio of heterozygous SNPs in the 28 tumour SNP arrays and re-assessed the performances of exome-based methods only on clonal events. Indeed, we noticed substantial increase in sensitivity, while specificity remained unchanged (Figure 30 and Figure 31).

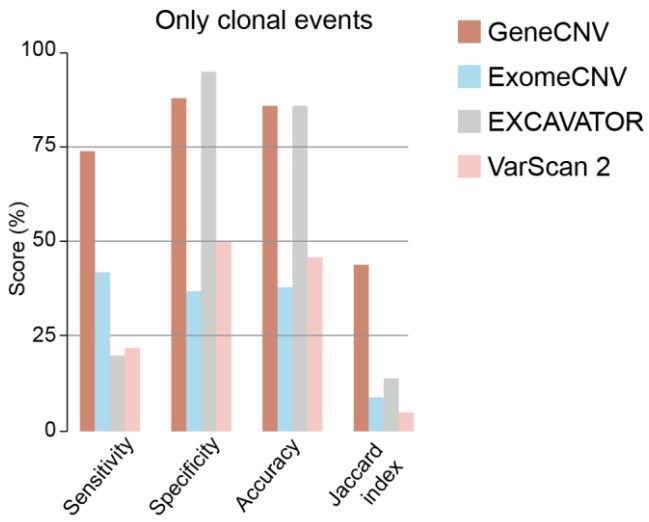


Figure 30: Comparison of GeneCNV with other exome-based methods for detection of clonal events in the 28 tumour exomes

Reported are sensitivity, specificity, accuracy and Jaccard index of the four exome-based methods (GeneCNV, ExomeCNV, VarScan 2, and EXCAVATOR) in detecting clonal variant genes in the 28 samples as compared to SNP array.

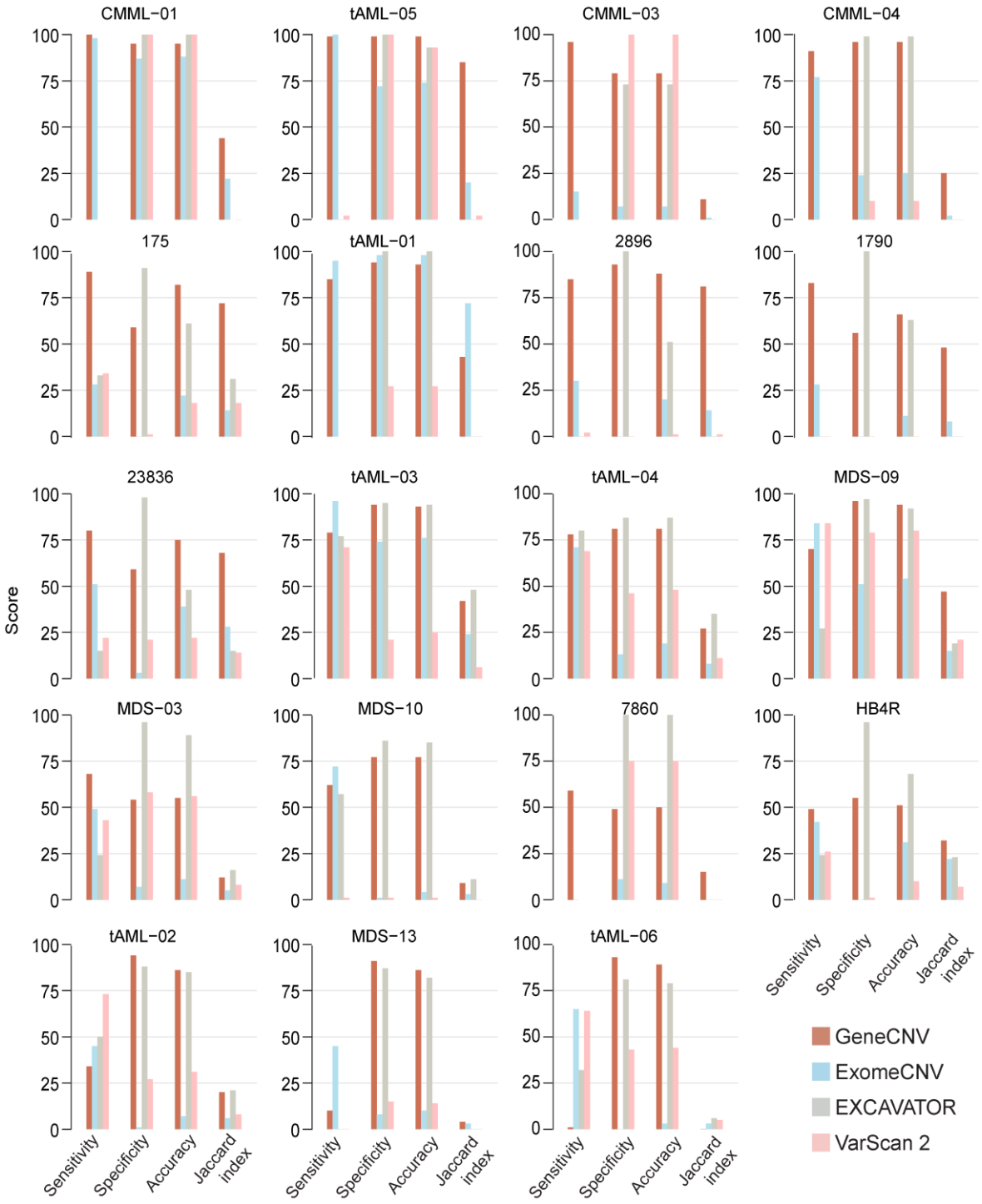


Figure 31: Performance assessment of the four exome-based methods in detecting clonal variants in each 28 samples

Reported are sensitivity, specificity, accuracy and Jaccard index of the four exome-based methods : GeneCNV, ExomeCNV, VarScan 2, and EXCAVATOR, in detecting clonal variant genes in each sample as compared to SNP array. Samples with at least one clonal variant are shown.

3.1.3 Analytical and graphical outputs

GeneCNV, in addition to providing the list of amplified, deleted and CN-LOH genes, also generates a graphical output with an analytical report of the results for each sample. This report includes information on the variation of gene coverage in the tumour and matched normal before and after normalization, which can be used to evaluate the coverage differences between the two exomes (Figure 32).

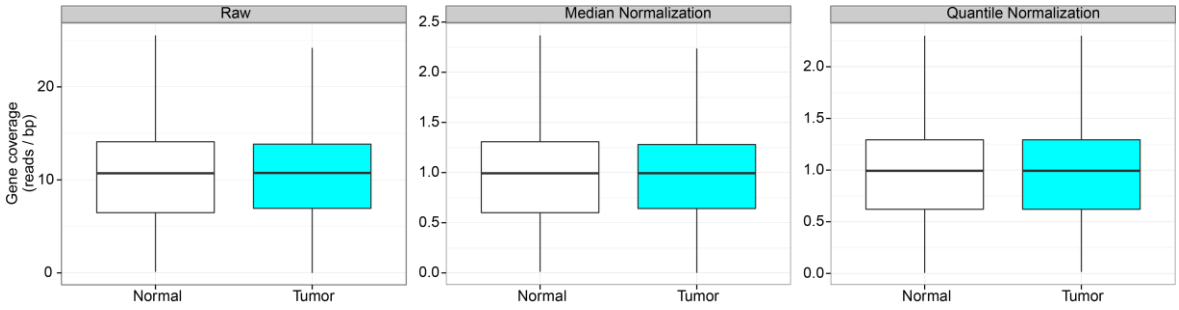


Figure 32: Distribution of genes coverage before and after normalization

Shown are the plots of gene coverage in the tumour and the normal exomes before and after the two normalizations: median (within sample) and quantile (between samples).

GeneCNV next provides the $L2R_{GC}$ spectrum of all targeted genes in the tumour exome, with amplified, deleted, and CN-LOH genes depicted in green, red, and yellow, respectively (Figure 33).

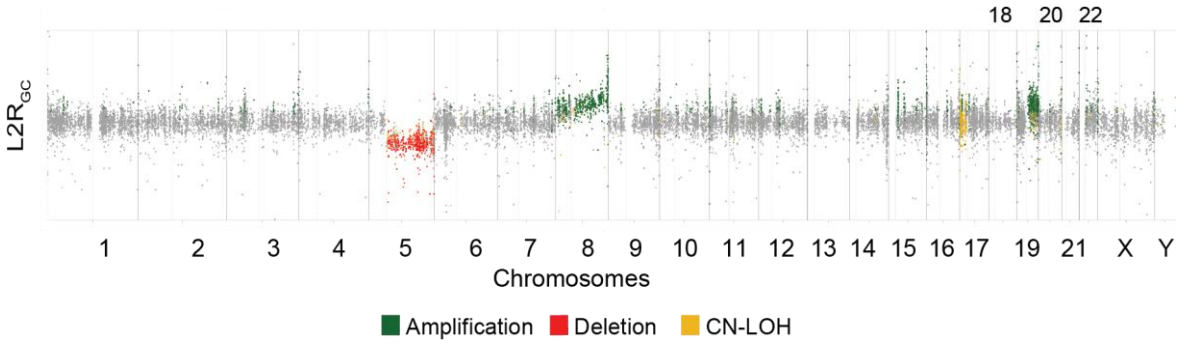


Figure 33: Spectrum of $L2R_{GC}$ in the tumour exome

Displayed are the tumour $L2R_{GC}$ spectrum with amplified, deleted and CN-LOH genes highlighted

It then provides a quantification of tumour-altered genes and the cumulative density map of their distribution along the chromosomes useful for understanding the most prevalent type of alteration in the sample (Figure 34). Finally, it summarizes all results in a Circos plot where altered genes that play known (Futreal, et al. 2004) or candidate (An, et al. 2014) roles in cancer are shown (Figure 35).

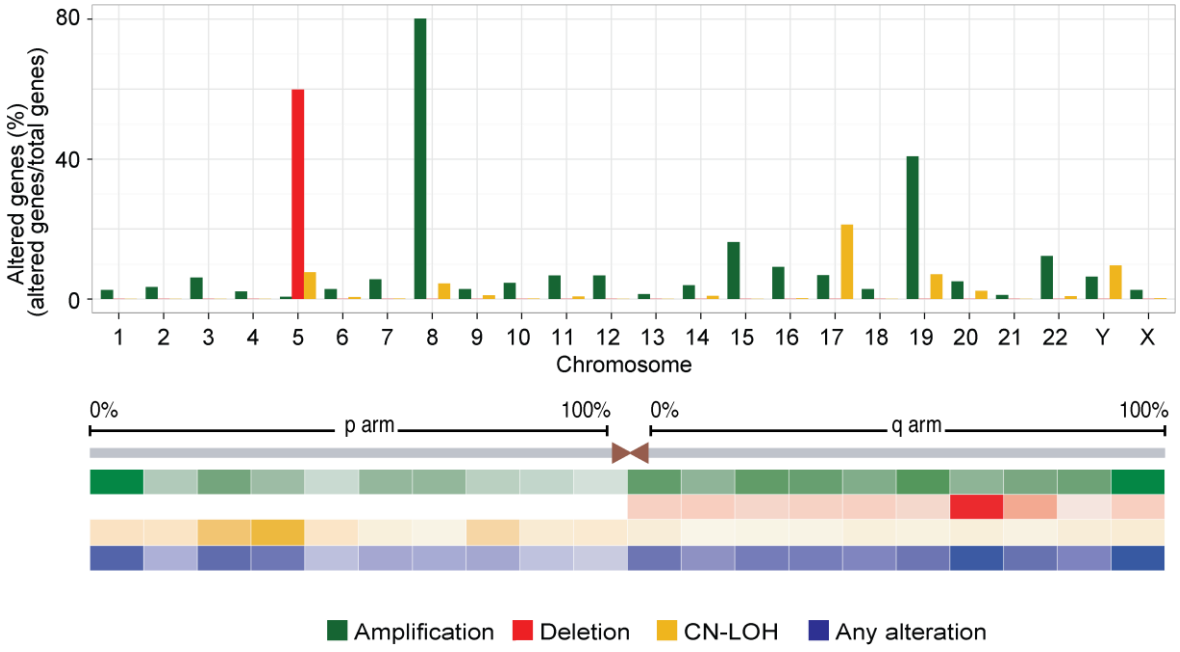


Figure 34: Distribution of altered genes in the sample

Shown is the percentage of somatic amplified, deleted and CN-LOH genes in each chromosome and in regions representing 10% of chromosome arms of the tumour exome

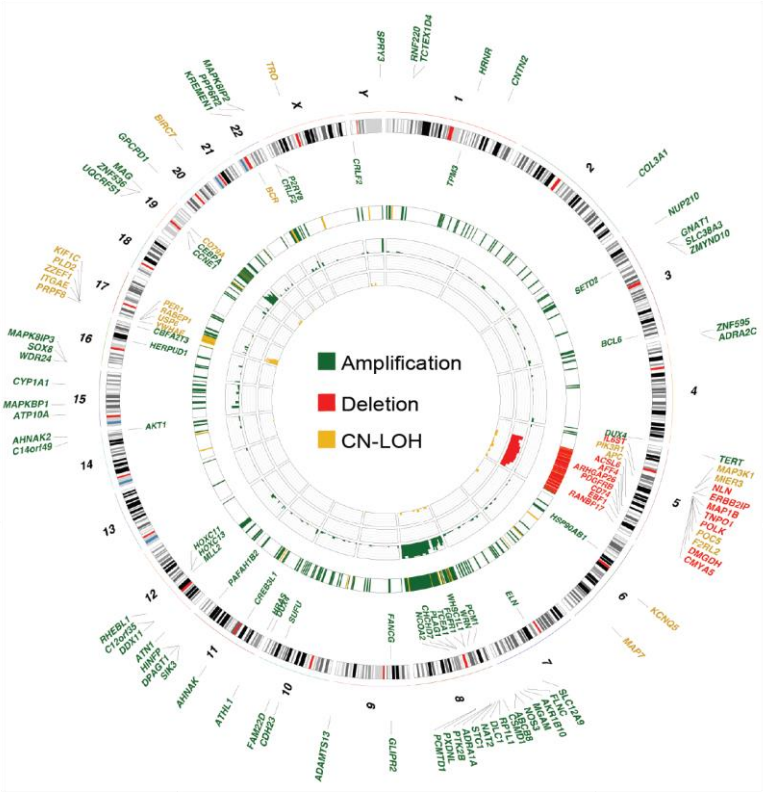


Figure 35: Circos plot summarizing the CNVs analysis for the sample

Shown is the Circos plot summarizing all alterations and reporting known and candidate cancer genes that undergo modifications in the tumour exome.

Apart from analysing tumour-normal exome pairs for CNVs and CN-LOHs, GeneCNV can also be used to compare CNV profiles across multiple samples. Such analysis is particularly useful when a cohort of patients is screened. In this case, the analysis report summarizes the percentage of total altered genes in each tumour exome (Figure 36A) and highlights the frequently altered genes across all samples (Figure 36B). The later analysis is particularly useful to pinpoint possible cancer driver events. It also provides the corresponding Circos plot and reports the recurrently altered cancer genes (Figure 37).

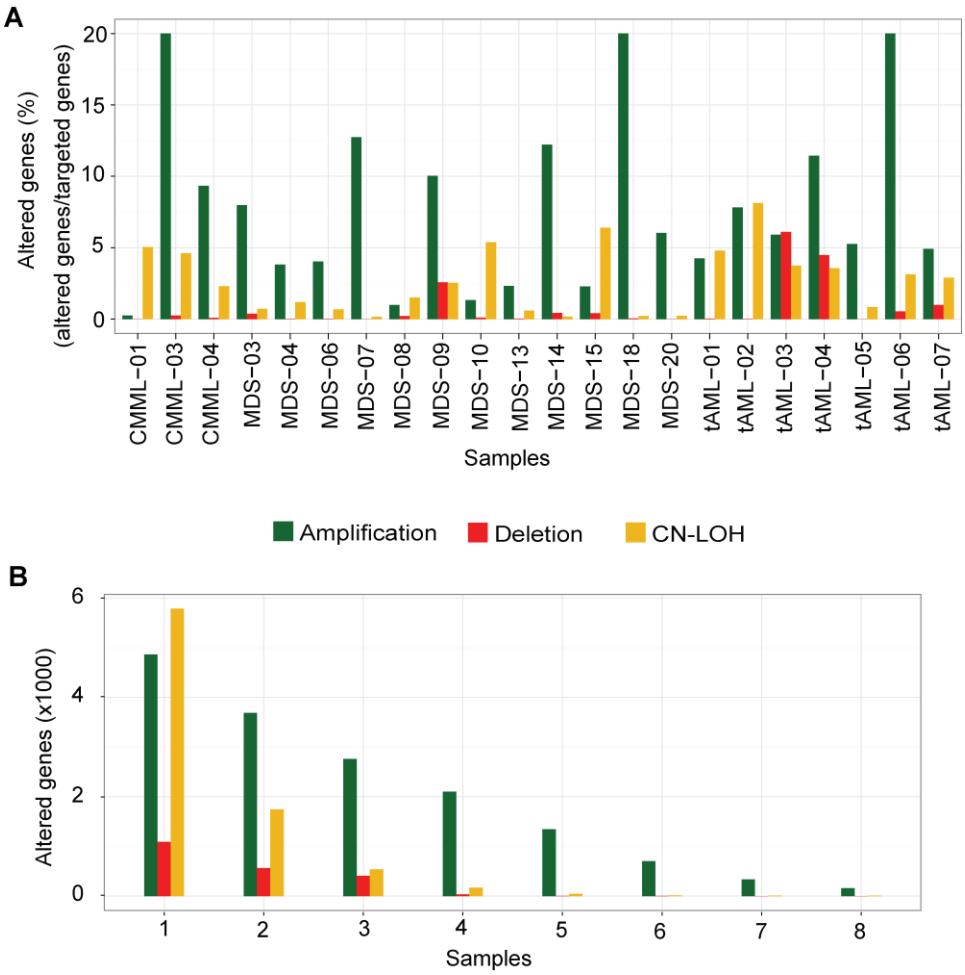


Figure 36: Distribution of amplified, deleted and CN-LOH genes across a cohort of tumour samples

Reported is (A) the percentage of amplified, deleted and CN-LOH genes in each tumour exome and (B) the number of genes whose modifications recur across samples present in the cohort

Figure 37: Circos plot with overview of the CNVs detected in the cohort of tumour samples

Shown is the Circos plot reporting all genomic alterations in each sample and recurrently altered cancer genes.

3.2 Genomic alterations in human BSEP-HCCs and mouse *Mdr2*-KO HCCs

After the comprehensive evaluation of GeneCNV, we next analysed the genomic landscapes in liver cancers.

Human BSEP-HCC arises in a background of fibrosis and chronic inflammation in the absence of other mutagenic factors. Thus, it provides an opportunity to understand the contribution of chronic inflammation, cirrhosis and fibrosis to the acquired genomic alterations that trigger liver cancer. *Mdr2*-KO mouse model is aetiologically similar to BSEP-HCCs patients. *Mdr2*-KO mice have inactivating mutation of bile transporter gene *ABCB4.* Impairment of *ABCB4* causes cirrhosis, fibrosis and chronic inflammation in liver leading to cancer. Hence, *Mdr2*-KO mouse provides the opportunity to validate the

contribution of genomic modification in BSEP-HCCs. We profiled the genomes and the exomes. We then applied VarScan 2 to detect SNVs and indels in both human and mouse. To detect CNVs, we used ASCAT for analysing SNP arrays for BSEP-HCCs and CNVnator for analysing WGS data for *Mdr*2-KO HCCs. In addition, we applied GeneCNV on WES and targeted re-sequencing data of *Mdr2*-KO HCCs for detecting CNVs.

In the following paragraphs I will describe the observations on the genomic alterations that occur in seven BSEP-HCCs and fourteen *Mdr2*-KO HCCs.

3.2.1 Human BSEP-HCCs do not accumulate mutations in known cancer genes

We performed somatic mutations analysis on the whole exomes of six BSEP-HCCs and corresponding background livers. We identified a total of 44 single nucleotide variants (SNVs) and 8 small insertions and deletions (indels) (Table 14) with no mutation shared between any two lesions.

Table 14: Somatic mutations and copy number alterations in human BSEP-associated HCCs

ID	*Sex*	*Age* ***(Years)***	*Tumour content*	*Somatic SNVs*	*Non-silent SNVs*	*Somatic Indels*	*Amplified genes*	*Deleted genes*	*LOH genes*
175	M	1.6	90%	7	3	0	10,688	1	4,964
7860	F	2.6	90%	5	2	0	13,594	1,248	891
23836	M	1.3	90%	5	1	0	12,450	19	2,847
HB4R	F	8.6	70%	25	9	7	5,598	2,575	7,124
1790	M	11.7	60%	1	0	0	8,601	0	158
2896	F	1.3	50%	1	0	1	9,687	244	3,702
UKT	M	1.3	40%	NA	NA	NA	3,801	258	0
Total	**-**	**-**	**-**	**44**	**15**	**8**	**18,428**	**3,628**	**9,757**

Shown are the SNVs and CNVs detected in the BSEP-HCCs along with tumour content. Non-silent SNVs indicate mutations leading to protein modifications. None of the identified indels introduced a frameshift. Total refers to modifications found in at least one sample. The higher number of mutations found in sample HB4R compared to other samples probably reflects that this was a relapse and that the patient had been treated with chemotherapy.

Each exome showed on average 0.05 somatic mutations per Mbp (Table 14), which was lower than the average mutation frequency observed in other human HCCs and other paediatric cancers (Vogelstein, et al. 2013) (Figure 38). This is in contrast to the adult HCCs that have higher mutation frequency. To exclude the possibility that we did not detect mutations due to poor sensitivity, we also measured the false negative rate of the variant calling method at different percentages of tumour content. We estimated 100% sensitivity in detecting somatic variants in lesions with >40% tumour content (Table 9).

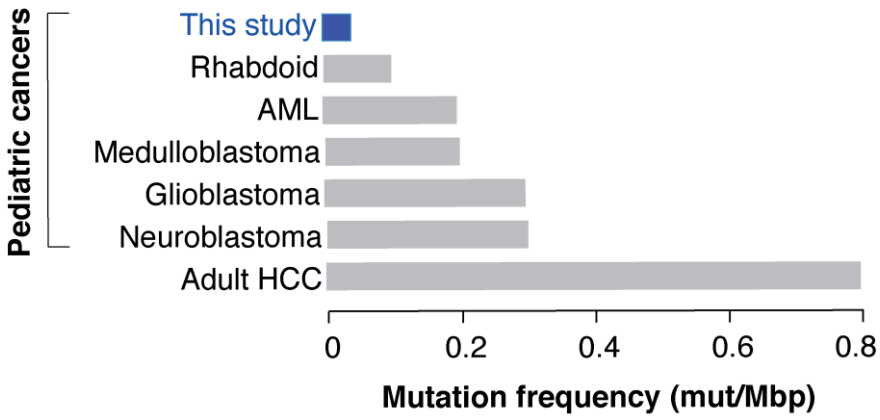


Figure 38: Mutation frequency in BSEP-HCCs, other liver cancer and paediatric cancers

Reported are the average non-silent mutations frequencies in the six BSEP-HCCs as paediatrics adult HCC and paediatric cancers (Vogelstein, et al. 2013).

We detected a total of 15 non-silent modifications in 15 genes (Table 15). These mutated genes were not known driver gene in HCC or in any other human cancers (Futreal, et al. 2004; Zhang 2012) for example *TP53*, *CTNNB1*, and chromatin regulators as mentioned previously (Figure 3). Also some were known passengers that recurrently mutate in several cancer types ((Lawrence, et al. 2013) and http://ncg.kcl.ac.uk/). Moreover, most of them were usually either poorly or not expressed in human liver thus indicating that they are probably not functional in this tissue (Table 15).

Overall, our results showed that BSEP-HCCs have low levels of mutational instability. The absence of mutations in known cancer driver genes suggests that the acquisition of mutational instability is not the driving force for the development of this type of liver cancer.

Table 15: Non-silent mutations in the seven human BSEP-HCCs

Chromosome	*Position*	*Reference*	*Variant*	*ID*	*Gene Name*	*RefSeq ID*	*Modification on cDNA*	*Modification on protein*	*Sanger validation*	*Expression in human liver*	*Recurrent passenger gene*
chr11	55541019	G	A	7860	*OR5D13*	NM_001001967	c.G106A	p.V36I	NA	-	YES
chr8	106815605	G	A	7860	*ZFPM2*	NM_012082	c.G3295A	p.E1099K	TRUE	No	No
chr6	152755062	C	G	23836	*SYNE1*	NM_182961	c.G4329C	p.M1443I	TRUE	No	YES
chr2	27354663	C	T	175	*PREB*	NM_013388	c.G1036A	p.V346M	TRUE	High	No
chr4	38051398	C	G	175	*TBC1D1*	NM_015173	c.A1789G	p.I597V	TRUE	Low	No
chr17	80391669	G	A	175	*HEXDC*	NM_173620	c.G418A	p.A140T	TRUE	-	No
chr5	140589586	G	C	HB4R	*PCDHB12*	NM_018932	c.G1107C	p.R369S	TRUE	No	No
chr6	138725657	C	T	HB4R	*HEBP2*	NM_014320	c.C25T	p.P9S	TRUE	High	No
chr7	17838695	C	G	HB4R	*SNX13*	NM_015132	c.G2381C	p.R794P	TRUE	Low	No
chr8	28574094	C	T	HB4R	*EXTL3*	NM_001440	c.C518T	p.A173V	TRUE	Low	No

Chromosome	*Position*	*Reference*	*Variant*	*ID*	*Gene Name*	*RefSeq ID*	*Modification on cDNA*	*Modification on protein*	*Sanger validation*	*Expression in human liver*	*Recurrent passenger gene*
chr9	71491651	G	A	HB4R	*PIP5K1B*	NM_003558	c.G259A	p.A87T	TRUE	Low	No
chr10	119134186	C	A	HB4R	*PDZD8*	NM_173791	c.G553T	p.A185S	TRUE	-	No
chr12	47629532	G	A	HB4R	*FAM113B*	NM_138371	c.G686A	p.G229E	TRUE	-	No
chr15	43531430	G	A	HB4R	*TGM5*	NM_004245	c.C790T	p.R264W	TRUE	Low	No
chr20	60942292	G	C	HB4R	*LAMA5*	NM_005560	c.C10G	p.R4G	NA	No	No

Shown are for each non-silent mutation shown are chromosome; genomic coordinate on the human genome (GRCh37/hg19); reference and variant base; number of reads for the position (total coverage) and for the variant base (variant coverage); lesion where the variant was found; gene name and RefSeq accession number; mutation effect at the nucleotide and protein level; outcome of Sanger validation. NA = the region could not be PCR amplified; expression levels in human liver; and if the gene has been reported as frequently mutated in other tumours. For the methodology used to assess expression levels and recurrent mutation see the Methods. - = data not available.

3.2.2 Massive gene amplification occurs in human BSEP-HCCs

We used genome-wide SNP arrays to investigate the occurrence of CNVs in seven BSEP-HCCs, including all tumours screened for point mutations and one additional lesion (Table 14). We detected a total of 18,428 and 3,628 genes that were amplified or deleted, respectively in at least one of the seven samples (Table 14 and Figure 39) with CNV frequency higher than that observed in adult HCCs (Figure 40). We also detected 9,757 genes with CN-LOH in at least one of the seven samples that recurrently affected chromosome 11 and to a lesser extent chromosome 16 (Figure 39).

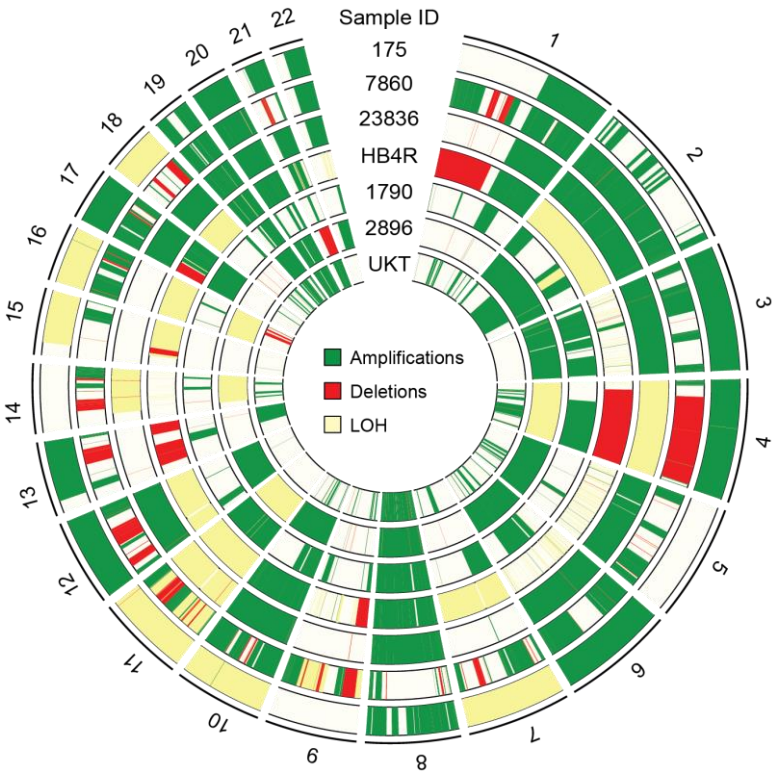


Figure 39: Copy number alterations in the 7 BSEP-HCCs

Circos plot reporting amplifications, deletions, and CN-LOH in all chromosomes of the seven BSEP-HCCs

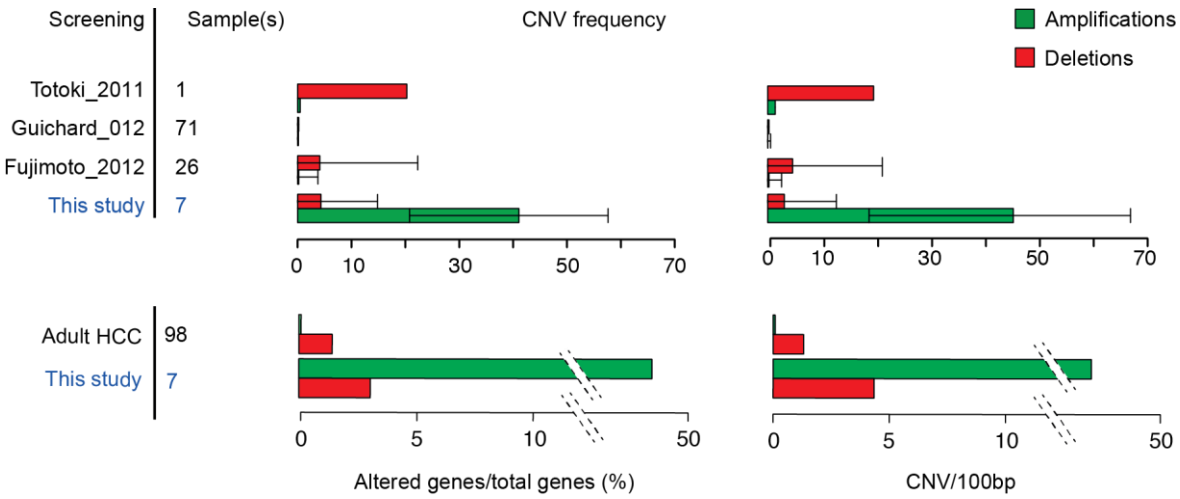


Figure 40: CNV frequency in BSEP-HCCs and adult HCCs

Shown is the frequency of amplifications (green) and deletions (red) in this study and other studies on adult HCCs. Above panel shows the CNV frequency in the seven BSEP-HCCs as compared to each adult HCCs (Totoki, et al. 2011; Fujimoto, et al. 2012; Guichard, et al. 2012). Below panel shows the average CNV frequency in the BSEP-HCCs and all adult HCCs take together. Frequency was measured as percentage of altered genes among total human genes and as CNVs per 100 base pairs. Bars indicate the minimum and maximum altered fractions across samples.

We found low-level copy number gains (on average 4 copies per aberrant region) as the most pervasive alterations, with around 38% of the total genome amplified in each sample (Figure 41A). We further analysed these samples for focal and arm-level CNV events (Beroukhim, et al. 2010) and observed that deletions were mostly focal and sample-specific with the exception of sample HB4R that had three arm-level deletions in chromosomes 1p, 4q and 17q (Figure 41). Instead copy number gains were arm-level or multiple focal events that led to the amplification of entire chromosomes, most notably in chromosomes 8, 19, and 20 in the majority of lesions (Figure 41B).

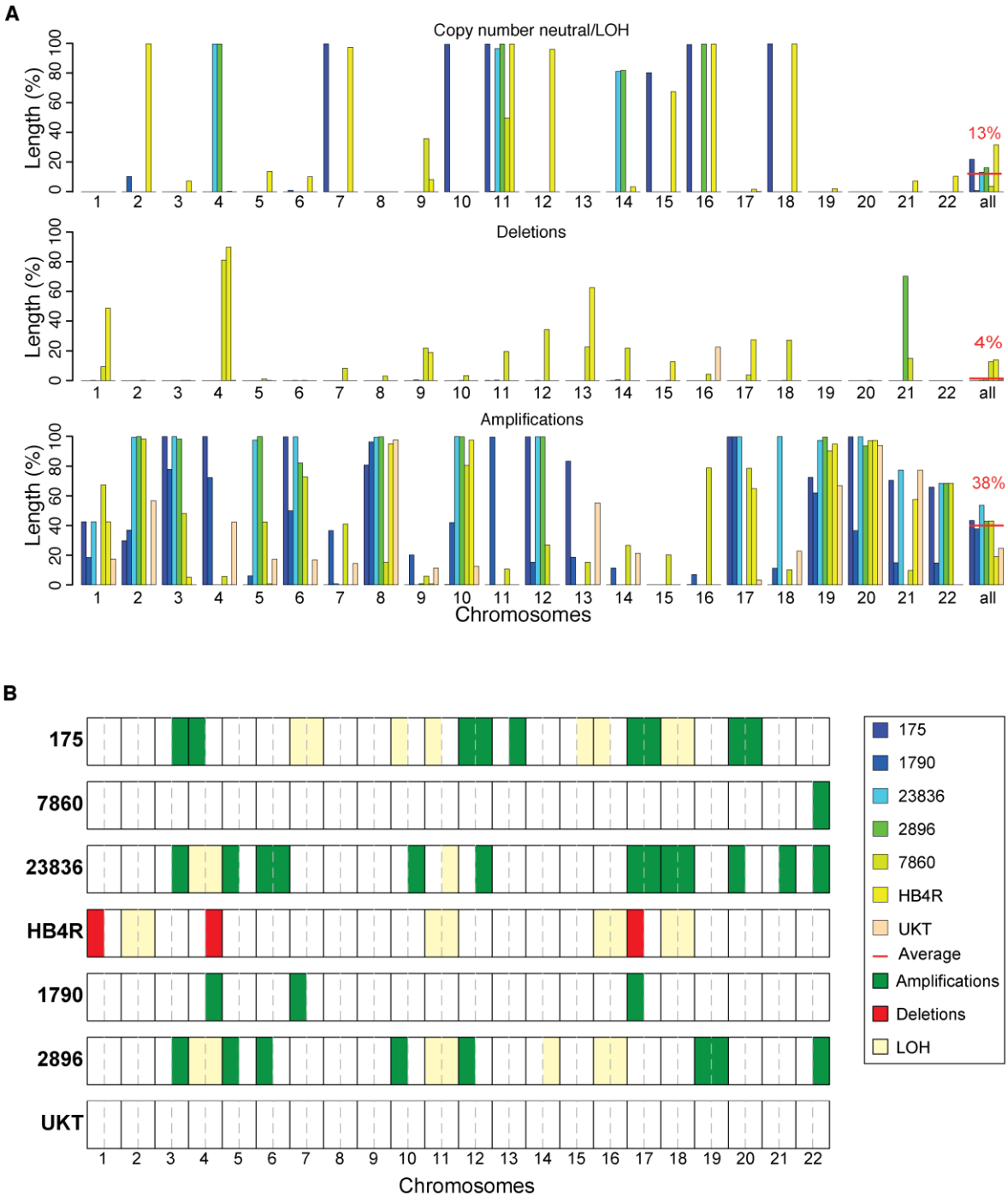


Figure 41: Distribution of genomic alterations in the 7 BSEP-HCCs

A) Percentage of chromosomes that undergo CN-LOH, deletions and amplifications in each samples. "all" refers to the average percentage of altered chromosomes for each sample and the percentage (red) is the average percentage across the seven samples. B) Arm-level CN-LOH events, deletions, and amplifications for each chromosome in each sample. Arm-level alterations were identified as single CNVs spanning ≥98% of the chromosome arm length.

To validate these results, we performed fluorescence in situ hybridization (FISH) on sample 23836 with probes located on both arms of chromosome 19 and confirmed the amplification of this chromosome (Figure 42).

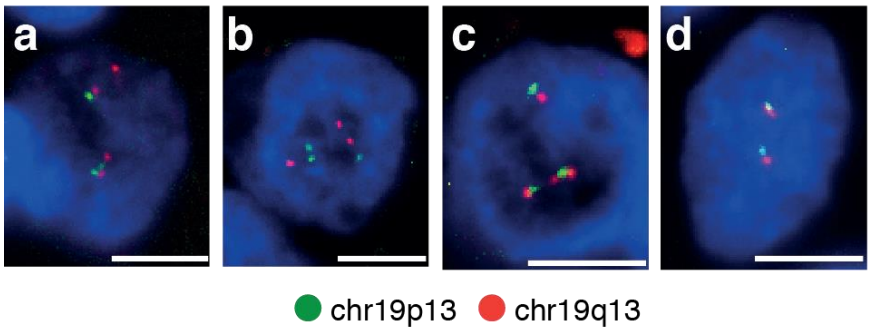


Figure 42: Amplification of chromosome 19 in the sample 23836 as validated by FISH

Shown are four representative hepatocyte nuclei (blue, DAPI staining) from a paraffin-embedded tumour sample 23836 confirming the amplification of chromosome 19 in patient 23836 as validated by FISH. Probes on chr19p13 (green) and on chr19q13 (orange) were used and nuclei in a-b-c show additional copies of chromosome 19, while the nucleus in d has only two. Scale bars = 20 um

Multiple amplifications likely occurred via step-wise DNA rearrangements rather than through one-shot catastrophic events, because we did not detect any sign of chromothripsis in most of the samples. The only exception was sample UKT (Figure 43), where we detected several consecutive oscillations between two copy number states in chromosomes 4 and 6, which may suggest the occurrence of catastrophic rearrangement events in this sample (Stephens, et al. 2011; Korbel and Campbell 2013).

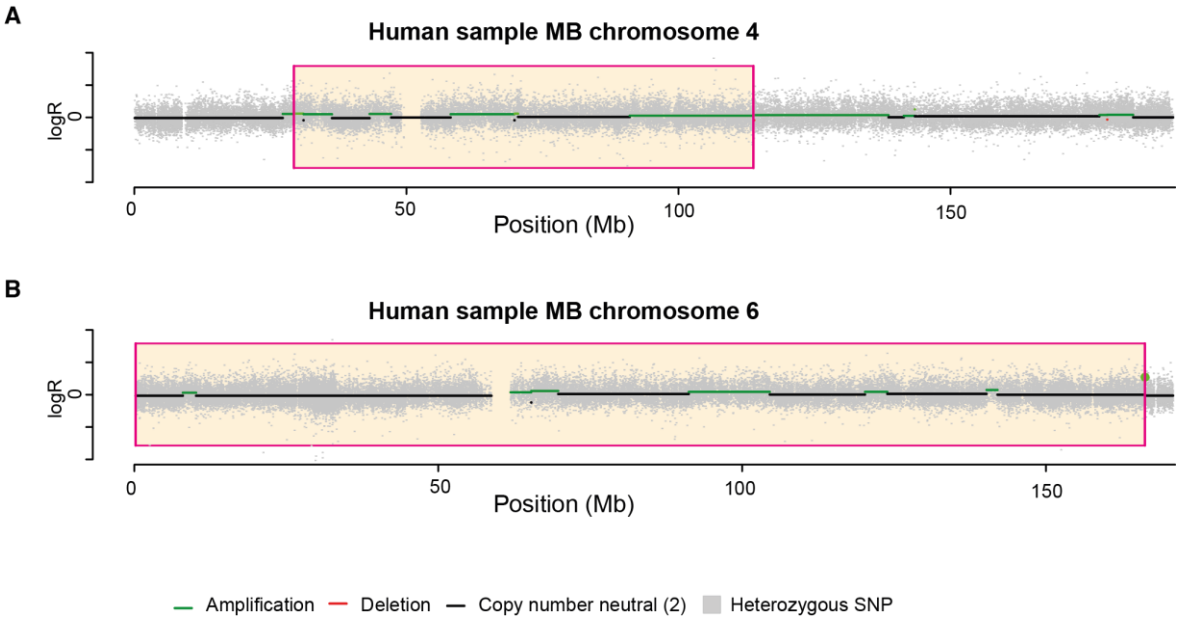


Figure 43: Oscillations of copy number alterations in sample UKT

Shown are the regions with at least ten oscillations between two copy number states are highlighted in pink for chromosomes 4 (A) and 6 (B) of human sample UKT Copy numbers were derived from ASCAT and adjacent regions with same copy number were merged before counting oscillations. Only the chromosome portion where ten consecutive changes between the same copy number states were observed is reported (Korbel and Campbell 2013).

To find possible cancer drivers we focussed on genes that were recurrently altered in the majority of samples and in particular on 935 known cancer genes that were amplified in the genome of at least four of the seven lesions (Figure 44). Pathway enrichment analysis of these genes highlighted three top-scoring pathways, namely the MAPK, the ErbB, and the PI3K/Akt pathways (corrected p-value =3x10-06, 9x10-06, and 2x10-05, respectively, hypergeometric test). These pathways form a complex and interconnected signalling network (Manning and Cantley 2007; Raman, et al. 2007) and their activation is a known driver event in some types of liver cancer (Calvisi, et al. 2006; Yea, et al. 2008; Calvisi, et al. 2011).

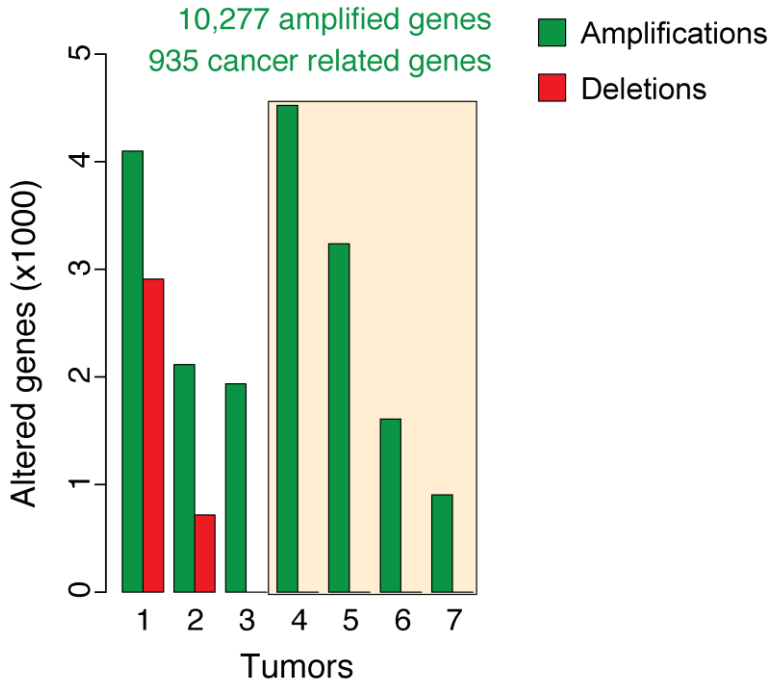


Figure 44: Recurrently altered genes in the 7 BSEP-HCCs

Reported are the amplified (green) and deleted (red) genes in the seven BSEP-HCCs, grouped according to the number of samples in which they were altered.

The results of CNV analysis showed that BSEP-HCCs are characterized by a pervasive occurrence of chromosomal rearrangements that lead to massive gene amplification. Recurrent events involve the amplifications of signalling genes, thus suggesting that the alteration of signalling pathways may be involved in the development and progression of HCC.

3.2.3 The genomic landscape of *Mdr2*-KO HCCs resembles that of BSEP-HCCs

To validate the potential contribution of the genomic alterations in BSEP-HCC, we profiled the cancer genomes of *Mdr2*-KO mice, which, like human BSEP-deficient patients, develop HCC due to impairment of bile formation (Smit, et al. 1993; Pikarsky, et al. 2004). We sequenced the exomes of nine HCCs extracted from the liver of seven *Mdr2*-KO mice using the kidney of one of them as a reference (Table 16) and identified a total of

118 somatic SNVs and no indels (Table 16). Also in this case, we confirmed >93% specificity of the variant calling. None of the 118 SNVs was shared between any two tumours, and 60 of them led to modifications in 60 proteins. As with BSEP-HCCs, no mutated gene was a known driver of HCC or of other cancers (Futreal, et al. 2004; Zhang 2012) and only a few were expressed in the liver (Su, et al. 2004) (Figure 45). We further sequenced the coding exons of 866 mouse orthologs of human cancer genes in four additional *Mdr2*-KO HCCs, using the normal liver as a reference. In this case, we increased the depth of sequencing coverage to further exclude that mutations might have been missed because of high intratumoral heterogeneity. Again we found no somatic mutations and no small indels in any cancer genes in any of the four samples.

Table 16: Somatic mutations and copy number alterations in *Mdr2*-KO HCCs

ID	*Sex*	*Age (Months)*	*Size* ***(cm)***	*Tumour content*	*Sequenced regions*	*Somatic SNVs*	*Non-silent SNVs*	*Amplified genes*	*Deleted genes*
51509/1	M	16	1.1	20%	Whole exome	8	5	59	0
60400/2	F	13	1.4	40%		8	3	113	0
218/1	M	15	1	50%		5	2	298	1
52686/1	F	15	0.7	50%		8	4	15*	2
58853/3	M	15	1.7	60%		20	8	631	0
60400/1	F	13	0.9	60%		9	3	455	0
58163/3	M	15	3	70%		17	6	333	1
58163/4	M	15	3	70%		39	27	625	0
215/1	M	14	1.8	80%		4	2	562	0
54913/10	F	10	0.1	NA	866 cancer genes	0	0	49	0
54913/8	F	10	0.5	NA		0	0	10	0
55481/10	F	10	0.3	NA		0	0	17	0
55484/4	F	10	3	30%		0	0	41	0
Total	-	-	-	-	-	**118**	**60**	**2,507**	**4**

Shown are the SNVs and CNVs detected in *Mdr2*-KO HCCs. Nodules with insufficient amount of tissue for histology inspection where defined as NA (not available). Non-silent SNVs indicate mutations leading to protein modifications. Total refers to modifications found in at least one sample. *TaqMan copy number assay assessed a high number of false negatives for this sample, thus suggesting an overall underestimation.

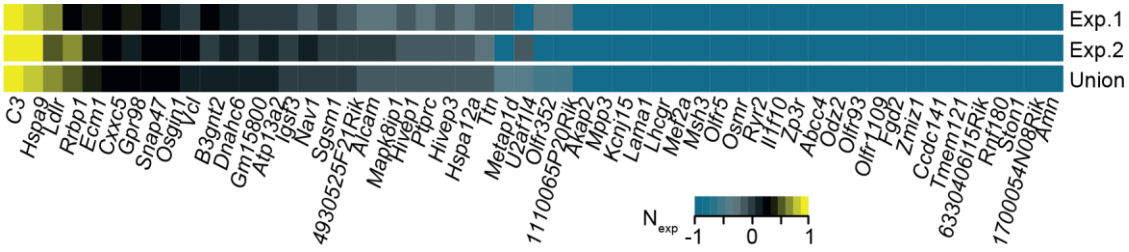


Figure 45: Expression levels of the mutated genes in normal liver of mouse

Shown are the expression levels of the mutated genes in normal liver (Su, et al. 2004) in the two replicates (Exp.1 and Exp.2) and in the average expression over the replicates (Union). Normalized gene expression (N_{exp}) was calculated as the gene expression level over the median expression of all genes in liver. Values above and below zero indicate gene expression higher and lower than the median. -1 indicates no expression in liver

To assess whether the massive copy number alteration observed in BSEP-HCC also occurred in mouse tumours, we used, GeneCNV, a novel method developed in our group. GeneCNV integrates frequency of heterozygous SNPs together with difference in the normalized gene coverage between the tumour and the reference; to detect CNVs directly from targeted sequencing data. In the 13 mouse HCCs, we identified a total of 2,510 altered genes, almost all of which were amplified (2,507, Table 16). We validated 10 randomly selected amplified genes with TaqMan copy number assay and estimated 70% sensitivity, 93% specificity and 84% accuracy of the method. To exclude the possibility that we did not detect deletions due to poor sensitivity, we compared the copy numbers of

genes on the male X chromosome with those of female samples in mouse and estimated 100% sensitivity and accuracy in calling deletions (Table 17). We further sequenced the whole genomes of two late stage *Mdr2*-KO HCCs from two different mice, using the corresponding kidneys as matched references. We again observed an overall higher occurrence of amplifications than deletions, with a total of 1,074 amplified and 117 deleted genes in the two genomes. For one of the two tumours (ID: 60400/1), which had been also used for exome sequencing, we observed that by far most of the amplified genes detected in the whole exome were also found in the whole genome (85%), thus further confirming the reliability of the method.

Table 17: Sensitivity assessment of the method in detecting deletions in *Mdr2*-KO HCCs

ID	*Sex*	*Reference ID*	*Sex*	*Genes on Chr X*	*Genes detected as deleted*	*Sensitivity (%)*	*Accuracy (%)*
218/1	M	52686/1	F	926	925	100	100
218/1	M	60400/1	F	926	926	100	100
218/1	M	60400/2	F	926	926	100	100

Shown are the sensitivity and accuracy in detecting deletions by comparing genes in chromosome X between male and female samples. Female samples are used as the reference sample. Reported are the sample ID and sex of the test and reference samples, genes present and deleted in chromosome X, sensitivity and accuracy. Sensitivity estimated as the number of genes detected as deleted in chromosome X over the total number of genes present in chromosome X.

The results of the genomic profiling of *Mdr2*-KO HCCs showed that, similarly to human BSEP-HCCs, these tumours are not prone to accumulate somatic point mutations

and small indels. Instead, they show pervasive chromosomal instability with overwhelming recurrence of gene amplification.

3.2.4 Somatic CNVs accumulate during *Mdr2*-KO HCC progression

Since BSEP-HCCs with high tumour content (90%) accumulated higher numbers of gene amplifications than those with low tumour content (Table 14), we used mouse HCCs to verify and quantify this signal. Interestingly, a positive correlation between tumour size and fold change of the amplified regions was already reported in HCCs from *Mdr2*-KO mice that underwent partial hepatectomy (Barash, et al. 2010). We confirmed a positive correlation between the number of amplified genes and the HCC content (Pearson correlation coefficient = 0.78, Figure 46). Moreover, bigger lesions showed significantly more amplified genes than smaller lesions (p = 0.03, Wilcoxon test, N = 10, Figure 46). This result suggests that amplifications tend to occur at later stages of tumour development. Furthermore, we also observed that amplifications accumulated preferentially near the centromeres (Figure 47) of mouse chromosomes, which are known hotspots for mitotic recombination (Jaco, et al. 2008).

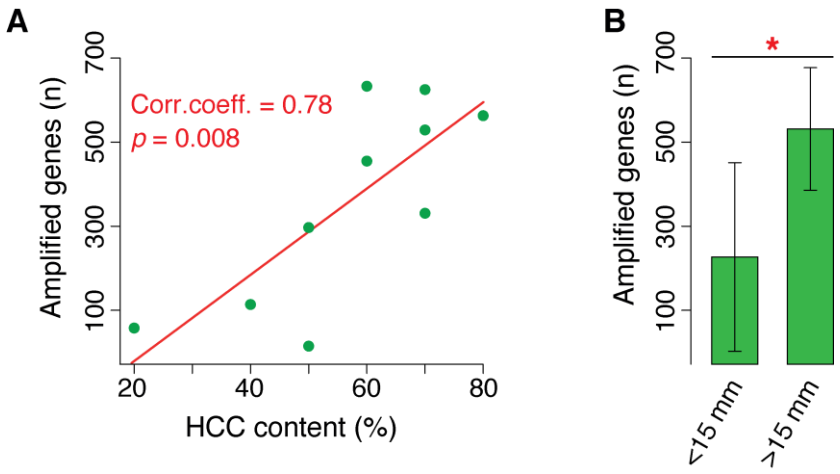


Figure 46: Accumulation of amplifications during tumour progression

Reported is the correlation between the number of amplified genes and HCC content in mouse tumours that underwent whole exome and whole genome sequencing (A) and the comparison

between number of altered genes in small (<15mm) and big (>15mm) lesions with minimum and maximum number of altered genes in the two groups are shown as bar (B). Correlation coefficients were measured using Pearson correlation testing. * = p <0.05, Wilcoxon test, N=10.

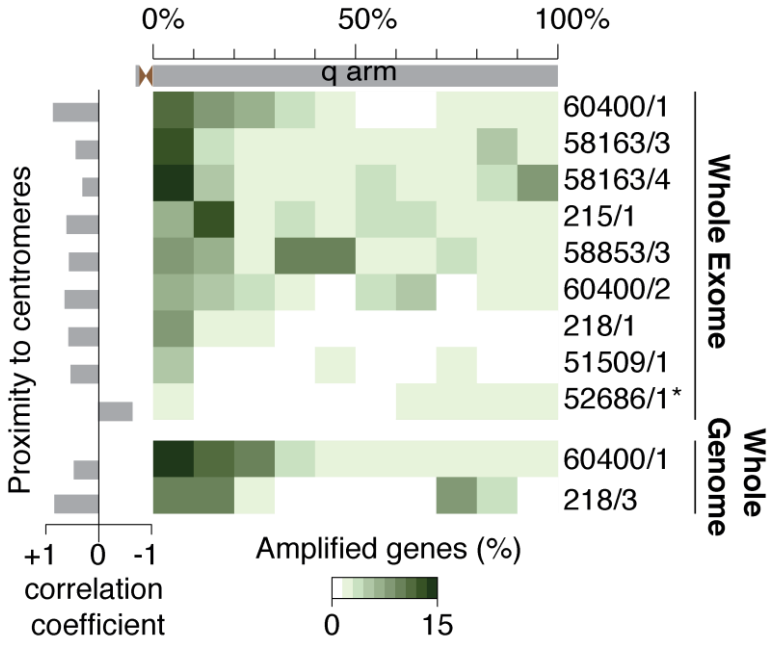


Figure 47: Preferential accumulation of amplifications near centromere

Shown are the cumulative fractions of gene amplifications (amplified genes/total genes) in regions representing 10% of the q arm length in all mouse chromosomes. In case of exome re-sequencing, the chromosome length was calculated as the region from the first to the last targeted base in the SureSelect XT Mouse All Exon kit (Agilent). Pearson correlation coefficients were calculated between the fraction of amplified genes in each region and the proximity to the centromere of each chromosome. *52686/1 was the only tumour with a negative correlation, likely due to overall underestimation of CNVs in this sample.

In both the *Mdr2*-KO HCC genomes we identified inverted translocations involving chromosomes 8 and 14 in one tumour (ID: 218/3, Figure 48A) and chromosomes 8 and 19 in the other (ID: 60400/1, Figure 48B). Interestingly, mouse chromosome 8 is the ortholog of human chromosomes 8 and 19, which are the most recurrently amplified

chromosomes in the human BSEP-HCCs. Through the analysis of discordantly aligned read pairs, we were able to map the two breakpoints at base pair resolution and both rearrangements were confirmed with PCR amplification and Sanger sequencing (Figure 48A and B). In one of the two tumours (ID: 60400/1), we counted ten consecutive disomic and trisomic copy number states in the region of chromosome 19 involved in the inverted translocation (Figure 48B and Figure 49). Therefore, again similarly to the human samples, we found possible indications that one-off catastrophic events could be responsible for the acquisition of at least some of the genomic rearrangements in this liver cancer type.

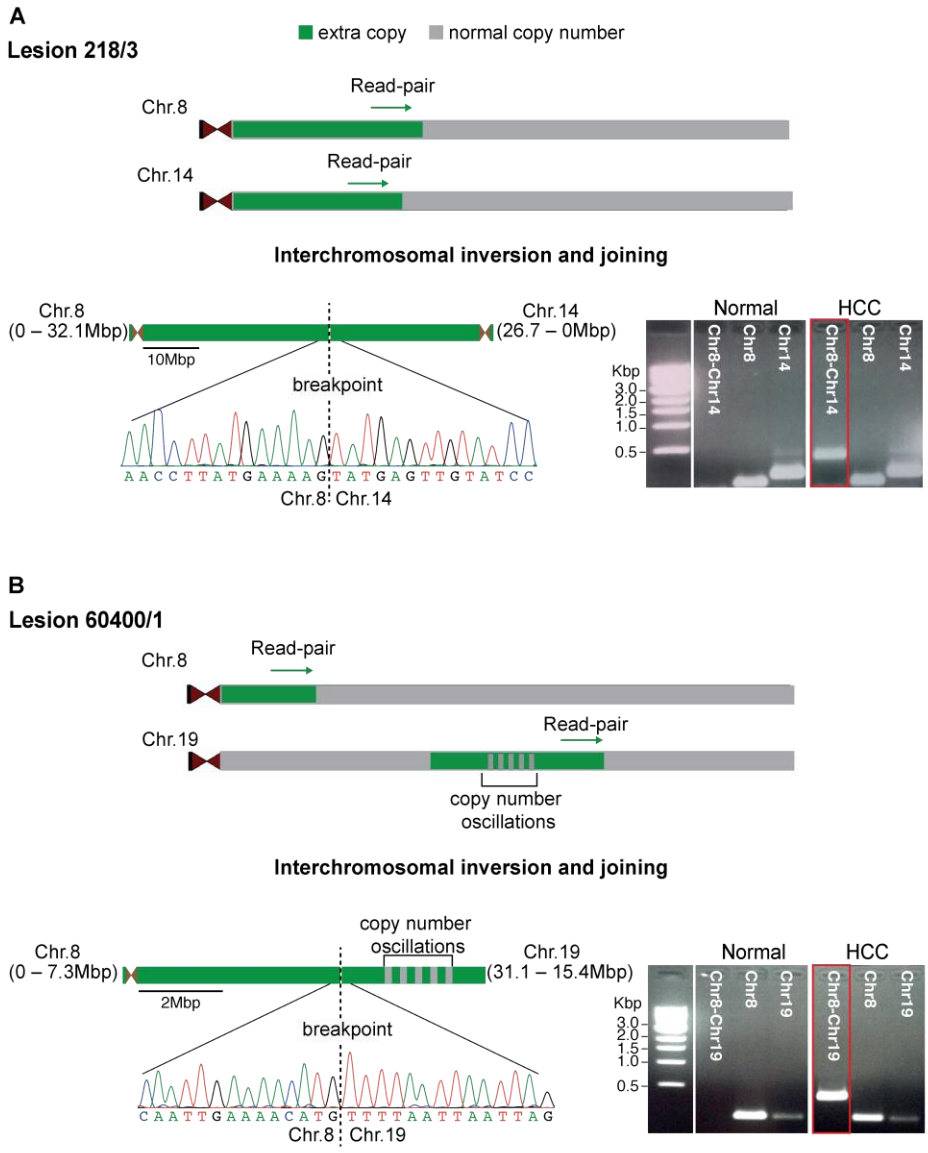


Figure 48: Inverted translocations in mouse samples 218/3 and 60400/1

Shown are the breakpoint locations at base resolution for inverted translocations between (A) chromosomes 8 and 14 of lesion 218/3 and (B) between chromosomes 8 and 19 of lesion 60400/1.

Through the analysis of discordant sequencing read pairs, breakpoints of both rearrangements were mapped at base pair resolutions and confirmed by PCR amplification and Sanger sequencing of breakpoint regions. In sample 60400/1, 10 consecutive copy number oscillations at chromosome 19 were detected in proximity of the rearrangement with chromosome 8. The genomic coordinates of amplified regions in each chromosome are shown in brackets.

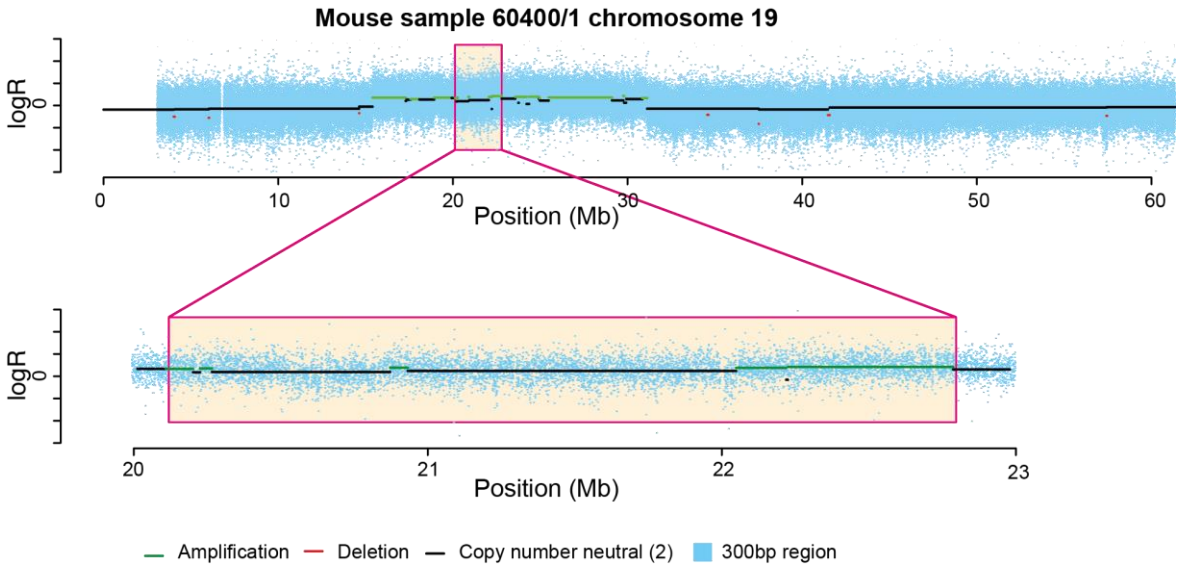


Figure 49: Oscillations of copy number alterations in Mdr2-KO HCC sample 60400/1

Shown is the region with at least ten oscillations between two copy number states is highlighted in pink for chromosomes 19 in mouse sample 60400/1. Copy number ratio for each region was calculated by dividing its normalized coverage by the normalized coverage of the normal counterpart over 300bp long regions. It should be noted that in case of the mouse sample 60400/1 the region of oscillation is probably longer, as shown by the whole chromosome diagram. We however reported only the chromosome portion where ten consecutive changes between the same copy number states were observed (Korbel and Campbell 2013).

In summary, the CNV analysis of *Mdr2*-KO HCCs showed that they undergo frequent genomic rearrangements that lead to gene amplifications. CNVs mostly occur at late stages of cancer development, thus suggesting a possible driver role in tumour progression.

3.2.5 JNK is deregulated in *Mdr2*-KO HCCs

Given the similarities in genomics landscape of acquired alterations between BSEP- and *Mdr2*-KO HCCs, we performed pathway enrichment analysis on 27 genes that were amplified in most human and mouse cancers (Figure 50). Again we found the MAPK signalling cascade among the top scoring pathways (corrected p = 2x10-02, hypergeometric test) (Table 18). In particular *Map2k7* was amplified in >70% human and mouse HCCs (5 out of 7 and 10 out of 14, respectively) (Figure 50).

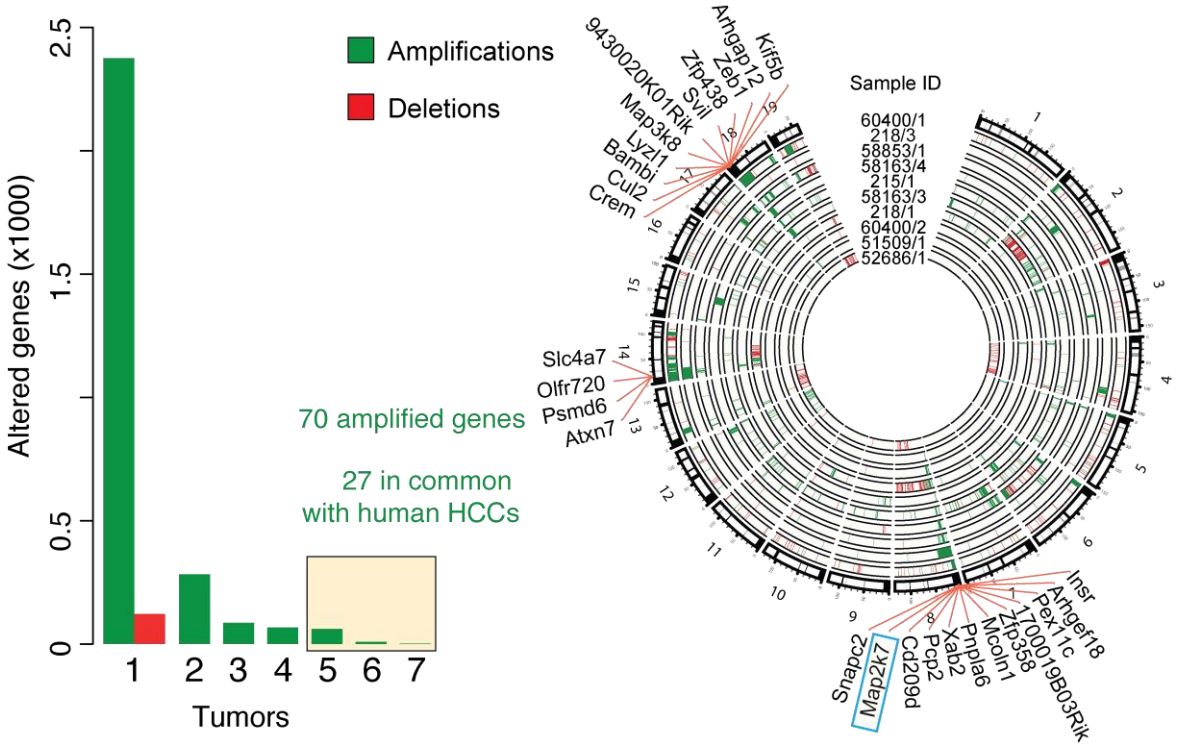


Figure 50: Recurrently altered genes in *Mdr2*-KO HCCs

Reported are the amplified (green) and deleted (red) genes in all the sequenced *Mdr2*-KO HCCs, grouped according to the number of samples in which they were altered. No gene was modified in more than seven tumours. Circos plot reporting amplifications and deletions in the 10 whole exome or whole genome sequenced samples with the 27 recurrently altered genes in both human and mouse. *Map2k7* was amplified in majority of the human and mouse HCCs (> 70%).

Table 18: List of enriched pathways in the 27 recurrently amplified genes in human and mouse tumours

Pathway	*Source*	*p-value*	*Corrected p-value*	*Amplified components*
Insulin Signalling	Wikipathways	0.0003	0.0115	*KIF5B; MAP3K8; MAP2K7; INSR*
MAPK signalling pathway	Biocarta/PID	0.0064	0.0230	*MAP3K8; MAP2K7*
Signalling by TGF-beta Receptor Complex	Reactome	0.0073	0.0230	*ARHGEF18; BAMBI*

Reported are the pathways found to be enriched in the 27 genes that are recurrently amplified in both human and mouse tumours (ConsensusPathDB (Kamburov, et al. 2013)), source from where the pathways information is taken, hypergeometric p-value, p-value after false discovery rate correction and amplified genes present in the pathway.

Since *Map2k7* was the most frequently amplified gene in *Mdr2*-KO HCCs, we screened 35 additional tumours from 16 distinct mice by TaqMan copy number assay to better quantify the frequency of its amplification. Altogether, *Map2k7* was amplified in 14 of the 49 *Mdr2*-KO nodules that were analysed overall (29%). This frequency was significantly higher than expected by chance (p=6x10-04, binomial test), but lower than the one we found previously in the exome sequenced samples. The difference was likely due to the presence of many adenomas among the screened samples. Indeed, 7 of the 12 nodules (58%) with high HCC content (≥40%) showed *Map2k7* amplification, a fraction significantly higher than expected by chance (p=9x10-05, binomial test) (Figure 51). This result further supports the general observation that gene amplification tends to occur preferentially in lesions with high tumour content.

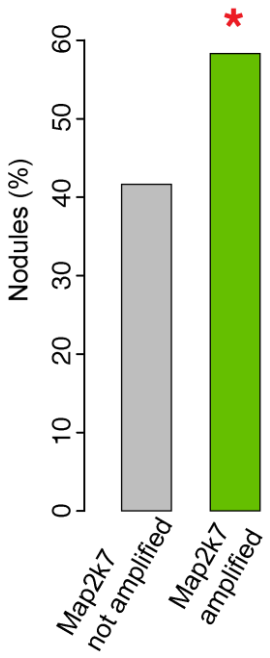


Figure 51: Map2k7 amplification in 12 *Mdr2*-KO samples with >40% HCC

Shown is the percentage of nodules with *Map2k7* amplification in the 12 *Mdr2*-KO tumour samples with >40% tumour content. * = p 9x10-05, Binomial test, N=12.

To investigate whether additional copies of *Map2k7* directly impinge on gene expression, we measured *Map2k7* mRNA levels in age-matched *Mdr2*-KO normal, inflamed, and tumoral liver samples. Significant overexpression was found in tumours with *Map2k7* amplifications compared to normal and inflamed *Mdr2*-KO liver (Figure 52) thus indicating that gene amplification led to increased expression.

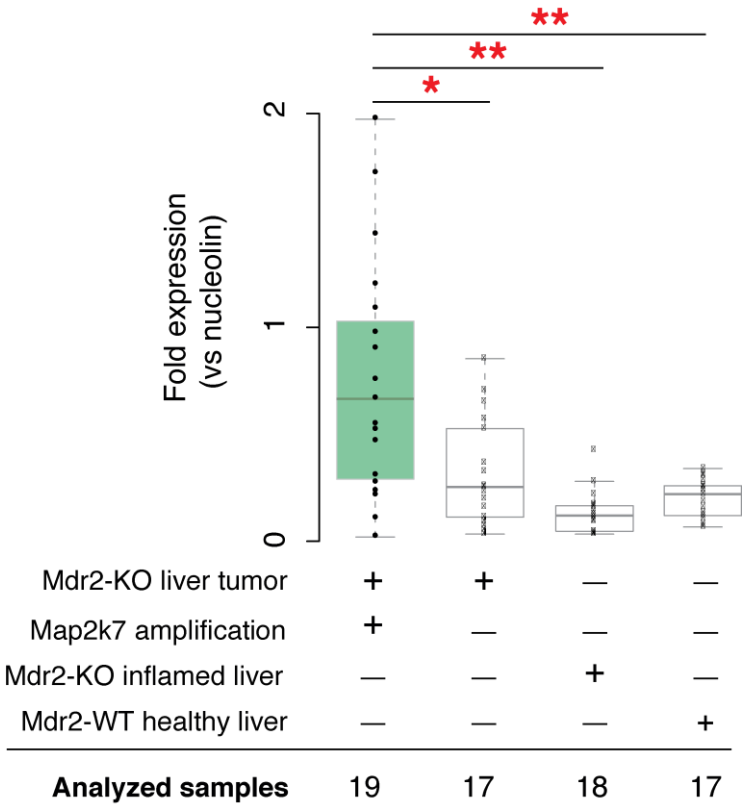


Figure 52: Map2k7 expressions in Mdr2-KO HCCs, inflamed and in normal samples

Shown is the *Map2k7* expression measured by qPCR in *Mdr2*-KO nodules where the gene is amplified, in nodules with no amplification, in *Mdr2*-KO inflamed livers, and in age-matched *Mdr2*-WT livers. Dots represent the different samples in each group. * = p <0.05, ** = p<0.005, Wilcoxon test, N=71. Maximum, minimum, and median are shown for each distribution.

Map2k7 is an upstream regulator of the c-Jun NH(2)-terminal kinase (JNK) (Davis 2000; Tournier, et al. 2001; Wada, et al. 2004), which is activated mainly by pro-inflammatory cytokines and environmental stress (Weston and Davis 2007). JNK deregulation has been already associated with liver cancer, where it plays cell- and stage-dependent roles during HCC development (Sakurai, et al. 2006; Hui, et al. 2007; Hui, et al. 2008; Beroukhim, et al. 2010; Nikolaou, et al. 2012). Interestingly, upstream and downstream direct JNK interactors, as well as JNK itself, were altered in several human and mouse samples (Figure 53).

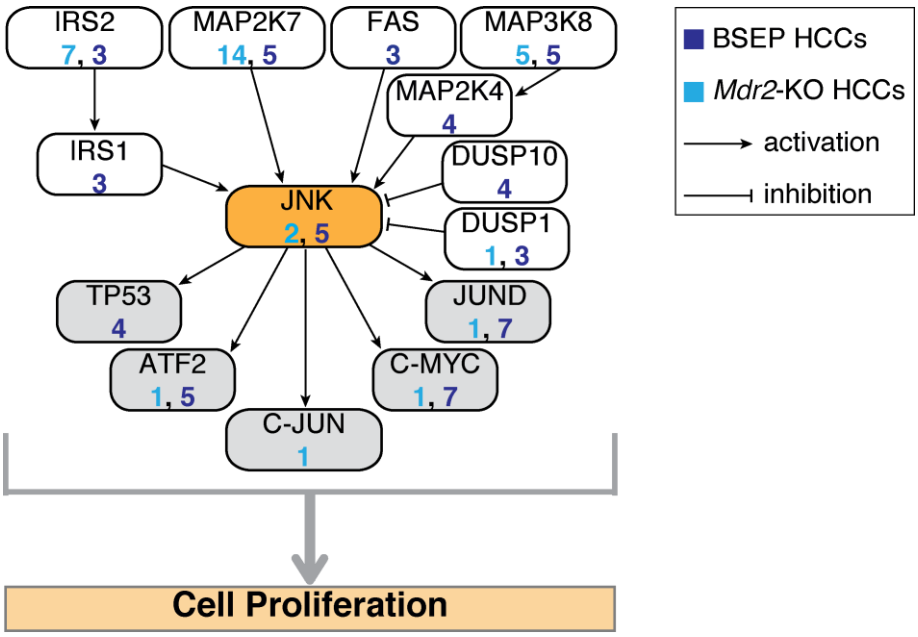


Figure 53: Frequent alteration of upstream and downstream direct JNK interactors in human and mouse HCCs

Shown are the 13 primary and two secondary JNK interactors that are frequently altered in *Mdr2*-KO and BSEP-HCCs. Numbers of lesions with the amplified gene are reported for mouse (light blue) and human (deep blue).

These data showed that gene amplifications occurring in *Mdr2*-KO tumours preferentially hit signalling genes, most notably JNK direct interactors, which may have a driver role in triggering liver tumour progression.

3.2.6 JNK inhibition arrests carcinoma progression in *Mdr2*-KO mice

We next investigated whether JNK inhibition might interfere with liver cancer progression by treating *Mdr2*-KO mice with SP600125, a synthetic polyaromatic chemical that directly inhibits the JNK kinases (Bennett, et al. 2001; Nikolaou, et al. 2012; Shibata, et al. 2012; Ye, et al. 2012; Lan, et al. 2013; Tan, et al. 2013). We randomized 23 *Mdr2*-

KO mice to receive either SP600125 or vehicle only (12 and 11 mice, respectively). We sacrificed mice after three weeks of treatment and compared the tumours from the two cohorts in terms of *Map2k7* amplification, nodule number, size, histology, and tumour content. We found a significantly lower proportion of lesions with *Map2k7* amplification in the treated mice (5 out of 36, 14%) in comparison to the untreated cohort (13 out of 35, 37%, p = 0.03, Fisher's exact test) (Figure 54). Thus, tumours with *Map2k7* amplification were more sensitive to JNK inhibition than those without *Map2k7* amplification, which explains their relative depletion after treatment. Indirectly, this result also indicated that the effects of SP600125 were mainly caused by its on-target activity on JNK. Despite of the two groups showing a comparable number of tumours per mouse, no mouse treated with the JNK inhibitor had nodules bigger than 20 mm, which instead represented ~20% of all lesions in the untreated group (Figure 55). In a reciprocal fashion, the proportion of nodules with diameters between 10 and 20 mm was significantly higher in treated mice than in the untreated group (Figure 55).

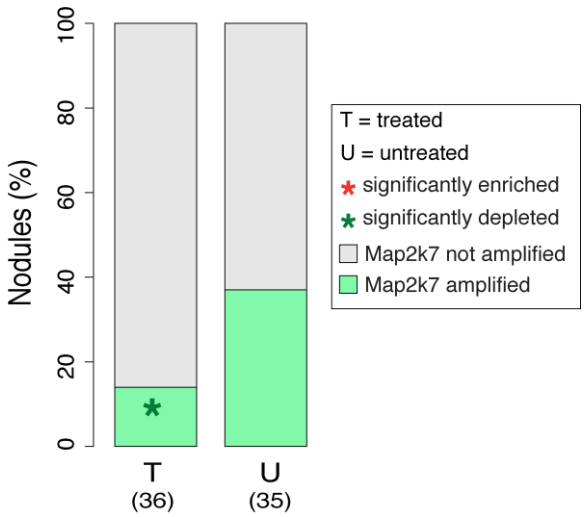


Figure 54: Nodules with *Map2k7* amplification

Reported is the percentage of nodules amplified in treated and untreated *Mdr2*-KO mouse groups. *Map2k7* is significantly amplified in less number of nodules in treated mouse group (p-value=0.03, Fisher's exact test)

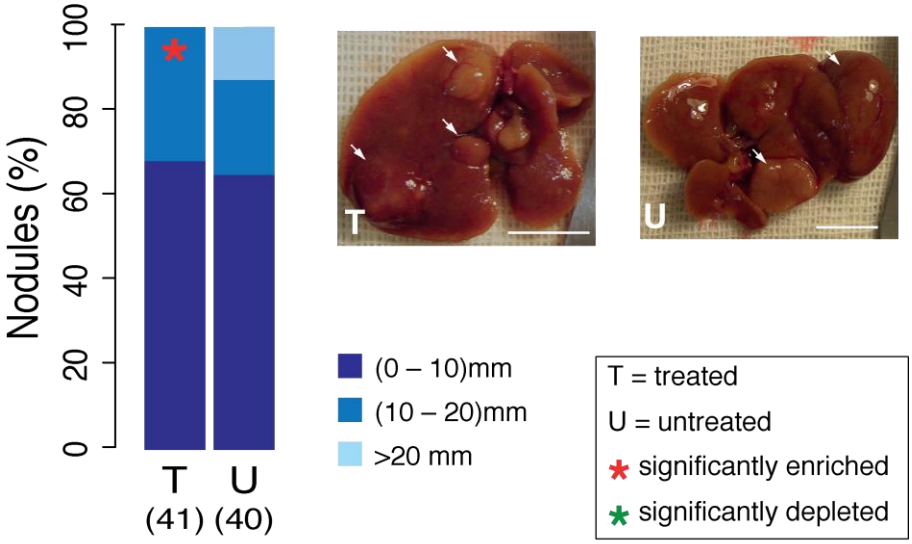


Figure 55: Nodule size in treated and untreated *Mdr2*-KO mouse groups

Reported are the size differences in nodules from treated and untreated *Mdr2*-KO mouse groups along with representative images of livers from a treated and an untreated mouse. Nodules from treated mice were significantly enriched in 10-20 mm lesions, but had no lesions >20 mm. The latter represented ~20% of nodules in untreated mice. Arrows indicate the nodules. Scale bar = 1cm

We then compared the histological composition of nodules in the treated and untreated groups and observed that nodules bigger than 10 mm showed significantly higher proportion of adenoma and lower proportion of adenocarcinoma in treated than in untreated mice (Figure 56). We did not detect any difference in the histological composition of small lesions (diameter <10 mm) between the two groups. Similarly, we compared the tumour content per mouse in the two cohorts, and again found that treated mice showed an overall significant depletion in HCC, while purely adenomatous nodules were over-represented (Figure 57)

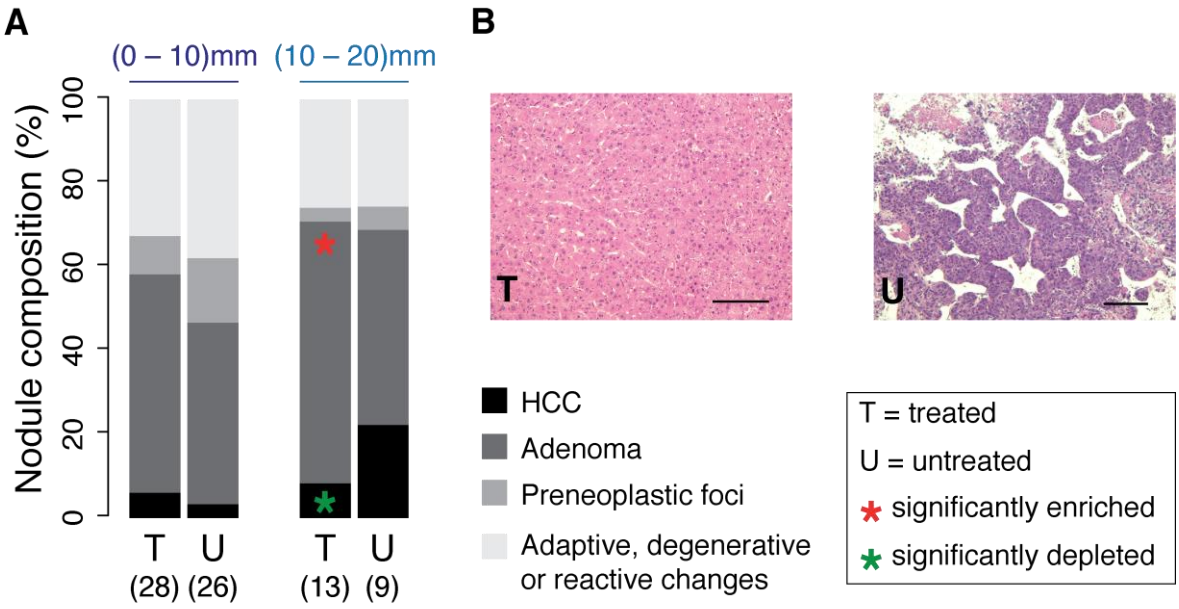


Figure 56: Histological composition of nodules in the cohort of treated and untreated mice

Reported is histological composition of nodules from treated and untreated mice along with representative photomicrograph histologic sections of HCC and adenoma from treated and untreated livers, respectively (hematoxylin / eosin). Nodules in the two cohorts of mice were divided into two groups by size (<10 mm and >10 mm) and the Scale bar = 150um.

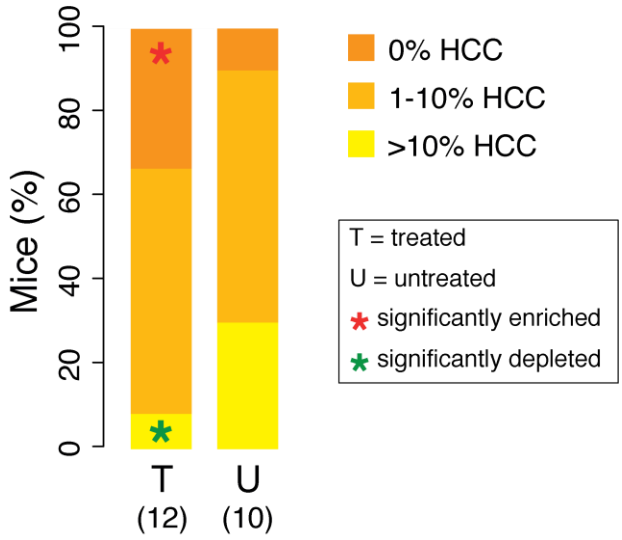


Figure 57: Tumour content in the treated and untreated mice

Shown is the cumulative tumour content in treated and untreated mice. Tumour content was measured as a percentage of HCC in each nodule. Nodules with HCC fraction >10%, <10%, and with no HCC were compared between treated and untreated mice.

Altogether, these data suggested that the drug blocks tumour progression towards bigger lesions with higher HCC content, thus supporting the role of JNK deregulation in progression more than in initiation of *Mdr2*-KO tumours. They were also consistent with the results of the CNV analysis, and specifically with the tendency of gene amplification, and of *Map2k7* amplification in particular, to occur in large lesions with high HCC content.

Discussion

HCC has a complex pathogenesis in which external factors such as virus infection, exposure to aflatoxin and alcohol all cooperate with inflammation to promote cancer. Recent studies have highlighted recurrent mutations in key cancer genes such as *TP53*, *CTNNB1*, and chromatin regulators and genomic regions that frequently undergoing CNVs (Li, Zhao, et al. 2011; Totoki, et al. 2011; Fujimoto, et al. 2012; Guichard, et al. 2012; Huang, et al. 2012; Jiang, et al. 2012; Sung, et al. 2012). Nevertheless, considerable differences in the acquired genomic alterations have been observed in these studies that suggest genetic heterogeneity of HCC depending on the initiating agents (Zhang 2012). So far, all the genomic re-sequencing studies in HCC have been mostly focussed on HCV- and HBV-associated HCCs (Li, Zhao, et al. 2011; Totoki, et al. 2011; Fujimoto, et al. 2012; Guichard, et al. 2012; Huang, et al. 2012; Jiang, et al. 2012; Sung, et al. 2012). However, HCCs invariably arise in the background of chronic inflammation, cirrhosis and fibrosis resulting from sustained liver injury. Currently, a clear understanding of how combination of chemical damage, inflammation and cirrhosis directly contribute to the acquisition of critical genomic changes that lead to cancer is lacking. Hence, we investigated the genomic landscapes of genomes of etiologically related human and mouse HCCs that develop cancer in response to chronic exposure to non-neutralized bile acids and in the absence of exogenous direct (viruses) or indirect (alcohol) mutagens. We studied the mutational and CNVs landscapes in both human and mouse HCCs. To this aim, we also developed a novel method, GeneCNV, for the detecting CNVs from WES data. In the successive paragraphs, I will first explain the rationale of GeneCNV and review its performance followed by detailed discussion on the identification and characterization of the genomic alterations in the human and mouse HCCs.

WES is one of the extensively used next generation sequencing technology for genotyping and identification of putative pathogenic mutations because it is time and cost effective. In addition, WES gives insights into the genomic alterations of protein coding

genes, which are often the most interesting to follow up. Although, a few methods have been developed to call CNVs from exome re-sequencing data in recent years, they all have some limitations (Liu, et al. 2013; Zhao, et al. 2013). First, all these methods use segmentation to identify CNVs. Ideally, segmentation will merge exons with same copy number into one segment such that the coverage variance is minimized within the same segment but maximized between adjacent segments (Figure 58). Segmentation has been successfully used in detection of CNVs from WGS and microarrays where information along the whole genome is available. However, its efficiency may be reduced when applied to WES data because of the scattered distribution of the exons in the genome. In particular, segmentations may lead to overestimation of CNVs in some cases (Figure 58). Second, none of the developed methods take into consideration the effect of normalization on samples undergoing large-scale rearrangement, which is a common feature of cancer. The distribution of $L2R_{GC}$ in such samples is skewed and the expected $L2R_{GC}$ for a diploid gene is shifted from the expected zero value. This may lead to ambiguous calls in methods such as VarScan 2 and EXCAVATOR that use fixed $L2R_{GC}$ values to define the thresholds for detecting altered genes. Another limitation of many methods is that they are unable to detect CN-LOH events from WES data. GeneCNV uses a novel approach to address these issues. First GeneCNV does not rely on segmentation of the genome to detect genomic alterations and instead calls copy number changes at the gene level by merging the targeted exons within a gene to form a contiguous representative gene. GeneCNV, then removes the bias in coverage due to sequence composition, DNA quality, library preparation protocols, and sequencing settings (Harismendy and Frazer 2009; Knierim, et al. 2011; Sims, et al. 2014) and make the gene coverages between the test and the reference samples comparable using normalization techniques. It next integrates the frequency of germline mutations (SNPs) with $L2R_{GC}$ to identify sample-specific thresholds for amplifications and deletions. Thus, the CNV calls made by GeneCNV are not affected by the modified spectrum of

$L2R_{GC}$ in samples with large-scale CNVs. In addition, GeneCNV is able to detect CN-LOH events by integrating SNP frequency with $L2R_{GC}$.

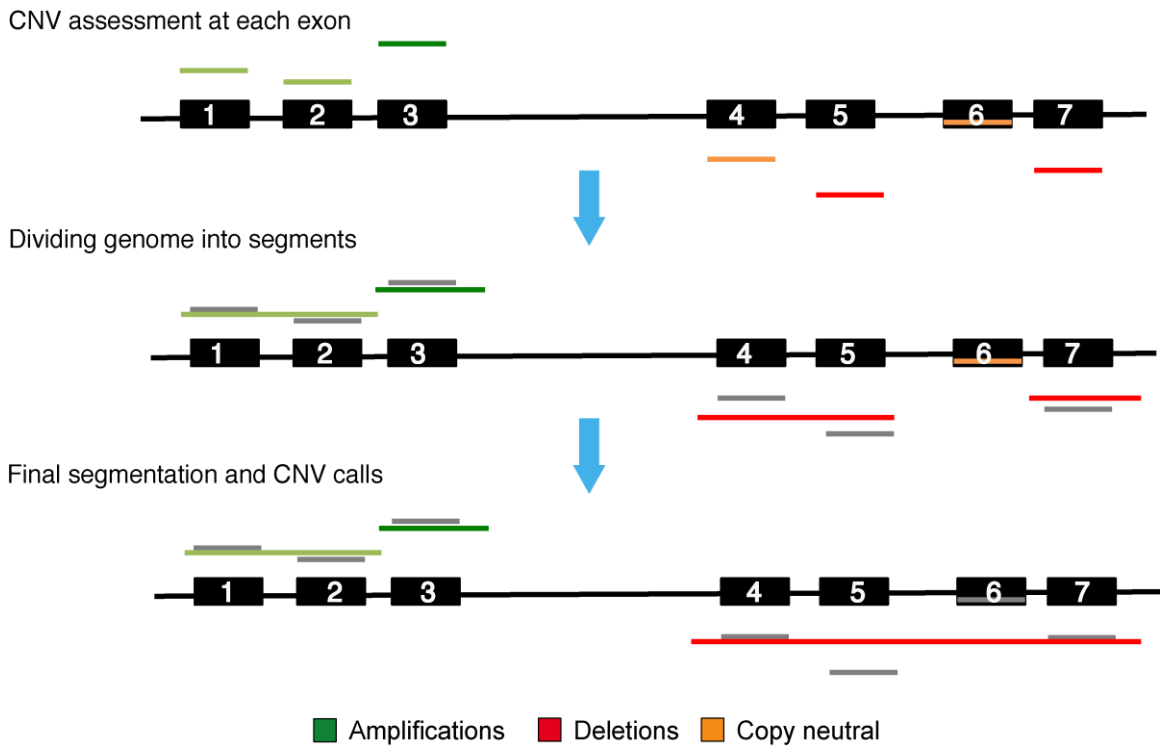


Figure 58: Segmentation approach for calling CNVs

Shown are the steps during segmentation and copy number call where calls at each exon are sequentially merged together with adjacent exons. Exons 1 and 2 (copy number=3) and exon 3 are amplified (copy number > 3). Exons 5 and 7 are deleted while exons 4 and 6 are copy-neutral. In the final CNV calls, copy neutral exons 4 and 6 are merged with deletion segments, which may or may not be true. In the situation that these are true, segmentation may lead to overestimation of deletions. Taken from (Sathirapongsasuti, et al. 2011)

We evaluated the performance of GeneCNV and compared with three other exome-based methods: ExomeCNV, VarScan2 and EXCAVATOR. In general, all exome-based methods showed low sensitivity in detecting CNVs. Nonetheless, GeneCNV performed better than the other existing methods and showed the highest concordance with the SNP

arrays (Figure 26). The poor sensitivity of all exome-based methods in majority of the samples were due to the subclonal genomic alterations that arise late in the cancer progression. We observed an increase in the sensitivity for all exome-based methods when only clonal variants were considered for the evaluation (Figure 30). Subclonal variants may be difficult to detect due to inappreciable differences in the coverage between tumour and normal samples, which may be further affected due to the non-contiguous nature of the exons. Though the current exome-based methods are limited to detecting clonal variants from WES data, GeneCNV performed better than other exome-based methods and had the highest sensitivity and concordance with SNP-derived CNV results. In addition, GeneCNV showed the highest sensitivity in detecting CN-LOHs (Figure 28). Altogether, GeneCNV provided the best solution in detecting CNVs and CN-LOHs among the exome-based methods studied here, thus highlighting the advantage of GeneCNV over other exome-based methods in detecting altered genes. Further improvements in GeneCNV can be anticipated by including information on tumour purity using a combinatorial approach integrating SNP frequency and $L2R_{GC}$ values. Since only limited combinations of mutation frequency and $L2R_{GC}$ values are possible for a given copy number state, any deviation from the expected values could be due to tumour heterogeneity or tumour purity (Liu, et al. 2013). This approach has been successfully applied in SNP arrays, where it has been shown to correctly identify the tumour purity and call copy number states (Attiyeh, et al. 2009; Popova, et al. 2009; Greenman, et al. 2010; Van Loo, et al. 2010; Yau, et al. 2010; Li, Liu, et al. 2011). After assessing the performance of GeneCNV in detecting altered genes, we next applied it to detect CNVs from exome of mouse HCCs.

We studied the landscape of somatic mutations and CNVs in seven human and fourteen mouse tumours. Despite the small number of samples in our study, the genomic profiling of both human and mouse showed a consistent genomic signature within and between species. The results of our screenings showed that this particular type of liver tumours accumulate a surprisingly low number of somatic SNVs and small indels. None of

the mutations detected were found in cancer genes such as *TP53*, *CTNNB1*, and chromatin regulators that were previously reported to play a driver role in HCCs (Li, Zhao, et al. 2011; Totoki, et al. 2011; Fujimoto, et al. 2012; Guichard, et al. 2012; Huang, et al. 2012; Jiang, et al. 2012; Sung, et al. 2012). Instead, these HCCs acquired massive CNVs with preferential accumulation of copy number gains (Figure 40). Our findings on *Mdr2*-KO tumours are in agreement with previous studies that have shown high degrees of chromosomal instability in *Mdr2*-KO HCCs with pervasive amplifications, no detectable deletions, and recurrent amplifications in chromosomes 5, 8, and 18 after partial hepatectomy (Barash, et al. 2010). In both human and mouse, we found possible indications that one-off catastrophic events could be responsible for the acquisition of at least some of the genomic rearrangements in this liver cancer type. The remarkable similarity in the genomic alterations between human and mouse HCCs suggests that *Mdr2*-KO mimics the molecular basis of this type of human liver cancer. Interestingly, none of analysed tumours showed a level of mutation and chromosomal instability that is comparable with other HCCs reported in recent studies (Li, Zhao, et al. 2011; Totoki, et al. 2011; Fujimoto, et al. 2012; Guichard, et al. 2012; Huang, et al. 2012; Jiang, et al. 2012; Sung, et al. 2012). These studies have mostly sequenced virus (HBV or HCV) or alcohol induced HCCs that acquire mutational instability and tend to accumulate gene deletions rather than amplifications (Guichard, et al. 2012). The genomic signature of HCC induced by bile acids and inflammation is unique and greatly differs from that of the other HCCs previously sequenced. These findings confirm the genetic heterogeneity of liver cancers caused by different etiological agents and, at the same time, the remarkable analogy among human and mouse tumours with similar etiopathogenesis.

We observed that both human BSEP-HCCs and mouse *Mdr2*-KO HCCs accumulate CNVs at late stages of cancer development. This may suggest a possible driver role of CNVs in tumour progression rather than initiation. The detailed analysis of CNVs in *Mdr2*-KO mouse cancer genomes revealed that copy number gains tend to cluster in

genomic regions with high mitotic recombination rates, such as centromeres and telomeres (Figure 47). Significant association has been observed for both CNVs and chromosomal rearrangements with increased recombination rate, hence making these regions as fragile sites in the genome (Volker, et al. 2010). Furthermore, these regions have been associated with LOH events (Gupta, et al. 1997) and replication stress (Dereli-Oz, et al. 2011; Burrell, McClelland, et al. 2013; Burrell, McGranahan, et al. 2013). Studies have shown that chronic inflammation induces the production of reactive oxygen species that may lead to oxidative DNA damage and reduced DNA repair causing genomic instability (Carter, et al. 2006; Barash, et al. 2010) (Figure 59). Moreover there is preliminary evidence for increased oxidative stress in the chronic inflammatory stages in *Mdr2*-KO mice (Katzenellenbogen, et al. 2006), which may explain the up-regulation of genes associated with chromosomal instability in the liver of *Mdr2*-KO mice (Carter, et al. 2006). It is therefore tempting to speculate that recombination hotspots are directly involved in cancer genomic instability in this tumour type, in the absence of external causes of DNA damage. This is also compatible with the role of inflammation in inducing a hypoxic microenvironment that favours chromosomal instability (Coquelle, et al. 1998; Eltzschig and Carmeliet 2011; Kumareswaran, et al. 2012). Inflammation induces hypoxia by increasing the metabolic demands of cells and reducing the metabolic substrates (Eltzschig and Carmeliet 2011). Hypoxia in turn induces breaks in fragile sites leading to gene amplifications (Coquelle, et al. 1998) through aberrant DNA double strand break repair (Kumareswaran, et al. 2012).

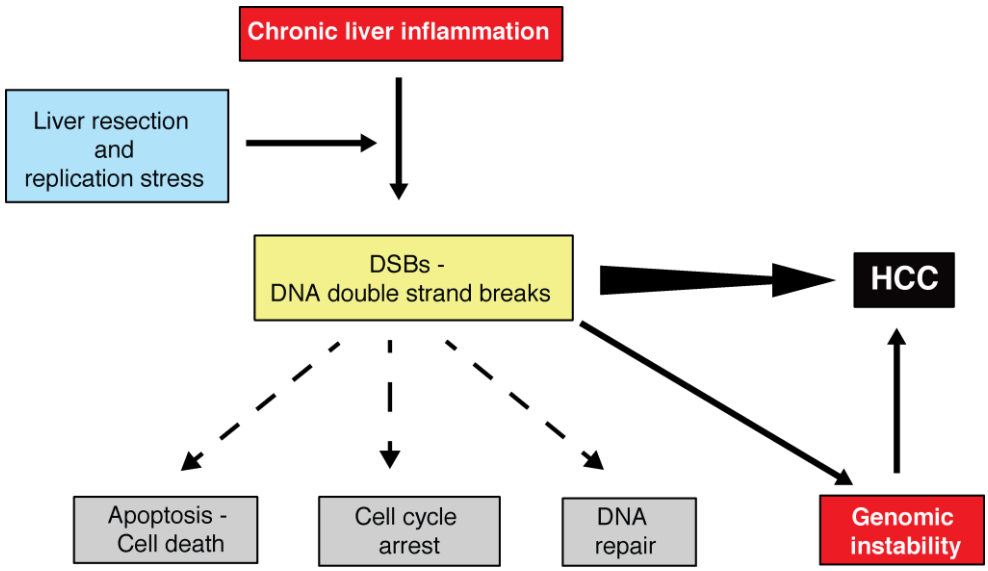


Figure 59: Chronic inflammation causes genomic instability leading to cancer

Chronic liver inflammation induces many afflictions, including DSBs, by oxidative damage. All these ailments, together and apart, contribute to the progress of HCC. The DNA damage response leads to cell cycle arrest, DNA repair, and apoptosis (dashed-line arrows). Accumulation of DNA damage results in genomic instability (solid-line arrow). Under the replicative stress, some of the impaired cells containing DSBs were salvaged from the DNA damage response and replicated, thus increasing genomic instability and facilitating tumour progression (solid-line arrow). Taken from (Barash, et al. 2010).

The analysis of recurrent copy number alterations in human and mouse HCCs was essential to pinpoint the similarities among tumours in the two species. In particular, we observed higher occurrence of amplifications than deletions. We observed MAPK signalling cascade among the top scoring pathways and identified recurrent amplifications of JNK activators. At least, one direct interactor of JNK or JNK itself was altered in all the samples and most notably *Map2k7* was amplified recurrently in both human and mouse HCCs. *Map2k7* are activated mainly by pro-inflammatory cytokines and environmental stress (Weston and Davis 2007). *Map2k7* were also overexpressed in mouse HCCs with *Map2k7* amplifications, thus suggesting the role of CNVs in its deregulation. Our findings

are also supported by previous studies in *Mdr2*-KO HCCs that have reported up-regulation of genes in the 20Mbps around the centromere of chromosome 8 where *Map2k7* is present (Katzenellenbogen, et al. 2007). These findings suggest that JNK pathway is deregulated and its modification may play a driver role in this liver cancer type. Since amplifications of JNK activators preferentially occur in the late stages of HCC development, the deregulation of this pathway is likely to favour cancer progression rather than initiation. JNK is involved in several physiological and pathological processes including cell proliferation, differentiation, apoptosis, and tumorigenesis (Weston and Davis 2007). Deregulation of JNKs has already been reported to play cell- and stage-dependent roles during HCC development and its activation has been implicated in the progression of liver cancer (Sakurai, et al. 2006; Hui, et al. 2007; Hui, et al. 2008; He, et al. 2010; Nikolaou, et al. 2012). For example, liver-specific deletion of *Mapk14*, a negative regulator of *Map2k7*, leads to JNK hyper-activation and HCC development in the mouse (Hui, et al. 2007). In addition, mice deficient in either *c-Jun* or *JNK1* do not develop HCC after being exposed to mutagens (Eferl and Wagner 2003; Sakurai, et al. 2006; Hui, et al. 2008; He, et al. 2010). Interestingly, JNK deficiency when present in both hepatocytes and non-parenchymal cells reduces the onset of inflammation and tumorigenesis. However its deficiency when limited to hepatocytes is linked to increased tumour size (Das, et al. 2011). This hints at an oncogenic role of JNK in non-parenchymal cells where it likely promotes an inflammatory environment that favours transformation and/or tumour progression. Our data support such a scenario, suggesting that JNK amplification leads to its deregulation and favours tumour progression in BSEP- and *Mdr2*-KO HCCs. In contrast, the JNK pathway is not recurrently amplified in virus-, alcohol- and other risk factor-associated HCCs (Li, Zhao, et al. 2011; Totoki, et al. 2011; Fujimoto, et al. 2012; Guichard, et al. 2012; Huang, et al. 2012; Jiang, et al. 2012). This further highlights that different disease aetiologies have distinct impacts on tumour genome, which in turn may lead to different molecular mechanisms of tumour initiation and progression.

In order to ascertain the role of JNK in HCC development, we inhibited JNK using the drug SP600125. Although SP600125 has been reported to exert secondary effects on targets other than JNKs (Dvorak, et al. 2008; Marozin, et al. 2012), we observed that tumours with *Map2k7* amplification were significantly depleted upon treatment indicating that the effects of SP600125 were mainly accounted for by the ability to inhibit JNK. No tumours from mice treated with the drug were of more than 20 mm in size, which instead represented ~20% of all lesions in the untreated mice. Moreover, tumours from treated mice showed significant depletion in HCCs and enrichment in adenomas when compared to tumours from untreated mice, thus suggesting that pharmacological inhibition of JNK impairs the adenoma-to-carcinoma progression *in vivo*. These finding pinpoint JNK as a potential target for therapy in this type of liver cancer and its inhibition may be useful to block HCC onset in BSEP deficiency patients waiting for liver transplantation.

In conclusion, the analysis of copy number alterations in liver cancers induced by chronic inflammation as a result of exposure of hepatocytes to bile acids, highlighted key genes and pathways that may play driver role in tumour progression. They also provide evidence that these tumours are associated with chromosomal instability rather than mutational instability. They demonstrate that these HCCs generate a unique and distinctive genomic signature that can be clearly distinguished from those caused by viruses and other external mutagens. They further confirm aetiology-dependent genetic heterogeneity in liver cancers. In the future, a comparative study of genomic features with tumours arising from other chronically inflamed tissues will be useful to determine the extent of similarity in the genomic changes as a characteristic of inflammation.

Appendix: Published papers

The two published papers and the manuscript under revision are attached at the end of the thesis. My contribution in each publication is reported below.

- Iannelli F[*], Collino A[*], Sinha S[*], Radaelli E, Nicoli P, D'Antiga L, Sonzogni A, Faivre J, Buendia MA, Sturm E, et al: **Massive gene amplification drives paediatric hepatocellular carcinoma caused by bile salt export pump deficiency.** *Nat Commun* 2014, **5:**3850. [*] *These authors contributed equally to this work.*

  Contribution: I had performed the CNV analysis from whole genome sequencing data and SNP array data for mouse and human samples respectively. In addition, I developed a novel method for detecting CNVs from targeted re-sequencing data, which was used in the identification of altered genes from whole exome sequencing data in the mouse samples. I had analysed the data from TaqMan copy number assay and quantification of *Map2k7* expression I had contributed in the writing of the manuscript.

- D'Antonio M, Pendino V, Sinha S, Ciccarelli FD: **Network of Cancer Genes (NCG 3.0): integration and analysis of genetic and network properties of cancer genes.** *Nucleic Acids Res* 2012, **40:**D978-983.

  Contribution: I had collected the human genes from Gencode and RefSeq and created a unique human gene set that was used in the protein network. I had helped in developing the web interface and contributed in the writing of the manuscript.

- Sinha S, Iannelli F, Gambardella G, Cereda M, Ciccarelli FD: **GeneCNV: detection of gene copy number variations and loss of heterozygosity from whole exome sequencing data.** *(Manuscript under revision).*

  Contribution: I had designed and developed the method. I had performed the CNV analysis from targeted re-sequencing data using the three exome-based methods (ExomeCNV, EXCAVATOR and VarScan 2) and SNP arrays, which is used as the gold standard for the assessment of the exome-based methods. I had contributed in the writing of the manuscript.

References

Abyzov A, Urban AE, Snyder M, Gerstein M. 2011. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res 21:974-984.

Aitman TJ, Dong R, Vyse TJ, Norsworthy PJ, Johnson MD, Smith J, Mangion J, Roberton-Lowe C, Marshall AJ, Petretto E, et al. 2006. Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. Nature 439:851-855.

Alison MR, Nicholson LJ, Lin WR. 2011. Chronic inflammation and hepatocellular carcinoma. Recent Results Cancer Res 185:135-148.

Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. Nat Rev Genet 12:363-376.

Alvarez L, Jara P, Sanchez-Sabate E, Hierro L, Larrauri J, Diaz MC, Camarena C, De la Vega A, Frauca E, Lopez-Collazo E, et al. 2004. Reduced hepatic expression of farnesoid X receptor in hereditary cholestasis associated to mutation in ATP8B1. Hum Mol Genet 13:2451-2460.

Amarasinghe KC, Li J, Halgamuge SK. 2013. CoNVEX: copy number variation estimation in exome sequencing data using HMM. BMC Bioinformatics 14 Suppl 2:S2.

An O, Pendino V, D'Antonio M, Ratti E, Gentilini M, Ciccarelli FD. 2014. NCG 4.0: the network of cancer genes in the era of massive mutational screenings of cancer genomes. Database (Oxford) 2014:bau015.

Arzumanyan A, Reis HM, Feitelson MA. 2013. Pathogenic mechanisms in HBV- and HCV-associated hepatocellular carcinoma. Nat Rev Cancer 13:123-135.

Attiyeh EF, Diskin SJ, Attiyeh MA, Mosse YP, Hou C, Jackson EM, Kim C, Glessner J, Hakonarson H, Biegel JA, et al. 2009. Genomic copy number determination in cancer cells from single nucleotide polymorphism microarrays based on quantitative genotyping corrected for aneuploidy. Genome Res 19:276-283.

Baca SC, Prandi D, Lawrence MS, Mosquera JM, Romanel A, Drier Y, Park K, Kitabayashi N, MacDonald TY, Ghandi M, et al. 2013. Punctuated evolution of prostate cancer genomes. Cell 153:666-677.

Balkwill F, Mantovani A. 2001. Inflammation and cancer: back to Virchow? Lancet 357:539-545.

Barash H, E RG, Edrei Y, Ella E, Israel A, Cohen I, Corchia N, Ben-Moshe T, Pappo O, Pikarsky E, et al. 2010. Accelerated carcinogenesis following liver regeneration is associated with chronic inflammation-induced double-strand DNA breaks. Proc Natl Acad Sci U S A 107:2207-2212.

Bassett AS, Marshall CR, Lionel AC, Chow EW, Scherer SW. 2008. Copy number variations and risk for schizophrenia in 22q11.2 deletion syndrome. Hum Mol Genet 17:4045-4053.

Bennett BL, Sasaki DT, Murray BW, O'Leary EC, Sakata ST, Xu W, Leisten JC, Motiwala A, Pierce S, Satoh Y, et al. 2001. SP600125, an anthrapyrazolone inhibitor of Jun N-terminal kinase. Proc Natl Acad Sci U S A 98:13681-13686.

Berasain C, Castillo J, Perugorria MJ, Latasa MU, Prieto J, Avila MA. 2009. Inflammation and liver cancer: new molecular links. Ann N Y Acad Sci 1155:206-221.

Bernstein H, Bernstein C, Payne CM, Dvorakova K, Garewal H. 2005. Bile acids as carcinogens in human gastrointestinal cancers. Mutat Res 589:47-65.

Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, Barretina J, Boehm JS, Dobson J, Urashima M, et al. 2010. The landscape of somatic copy-number alteration across human cancers. Nature 463:899-905.

Block TM, Mehta AS, Fimmel CJ, Jordan R. 2003. Molecular viral oncology of hepatocellular carcinoma. Oncogene 22:5093-5107.

Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, Janoueix-Lerosey I, Delattre O, Barillot E. 2012. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics 28:423-425.

Bolstad BM, Irizarry RA, Astrand M, Speed TP. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19:185-193.

Brechot C, Pourcel C, Louise A, Rain B, Tiollais P. 1980. Presence of integrated hepatitis B virus DNA sequences in cellular DNA of human hepatocellular carcinoma. Nature 286:533-535.

Breuhahn K, Longerich T, Schirmacher P. 2006. Dysregulation of growth factor signalling in human hepatocellular carcinoma. Oncogene 25:3787-3800.

Burrell RA, McClelland SE, Endesfelder D, Groth P, Weller MC, Shaikh N, Domingo E, Kanu N, Dewhurst SM, Gronroos E, et al. 2013. Replication stress links structural and numerical cancer chromosomal instability. Nature 494:492-496.

Burrell RA, McGranahan N, Bartek J, Swanton C. 2013. The causes and consequences of genetic heterogeneity in cancer evolution. Nature 501:338-345.

Cahan P, Li Y, Izumi M, Graubert TA. 2009. The impact of copy number variation on local gene expression in mouse hematopoietic stem and progenitor cells. Nat Genet 41:430-437.

Calvisi DF, Ladu S, Gorden A, Farina M, Conner EA, Lee JS, Factor VM, Thorgeirsson SS. 2006. Ubiquitous activation of Ras and Jak/Stat pathways in human HCC. Gastroenterology 130:1117-1128.

Calvisi DF, Wang C, Ho C, Ladu S, Lee SA, Mattu S, Destefanis G, Delogu S, Zimmermann A, Ericsson J, et al. 2011. Increased lipogenesis, induced by AKT-mTORC1-RPS6 signalling, promotes development of human hepatocellular carcinoma. Gastroenterology 140:1071-1083.

Cameron D, Casey M, Press M, Lindquist D, Pienkowski T, Romieu CG, Chan S, Jagiello-Gruszfeld A, Kaufman B, Crown J, et al. 2008. A phase III randomized comparison of lapatinib plus capecitabine versus capecitabine alone in women with advanced breast cancer that has progressed on trastuzumab: updated efficacy and biomarker analyses. Breast Cancer Res Treat 112:533-543.

Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, Stebbings LA, Leroy C, Edkins S, Hardy C, et al. 2008. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet 40:722-729.

Carter SL, Eklund AC, Kohane IS, Harris LN, Szallasi Z. 2006. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. Nat Genet 38:1043-1048.

Chen F, Ananthanarayanan M, Emre S, Neimark E, Bull LN, Knisely AS, Strautnieks SS, Thompson RJ, Magid MS, Gordon R, et al. 2004. Progressive familial intrahepatic cholestasis, type 1, is associated with decreased farnesoid X receptor activity. Gastroenterology 126:756-764.

Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. 2013. Emerging landscape of oncogenic signatures across human cancers. Nat Genet 45:1127-1133.

Clifford RJ, Zhang J, Meerzaman DM, Lyu MS, Hu Y, Cultraro CM, Finney RP, Kelley JM, Efroni S, Greenblum SI, et al. 2010. Genetic variations at loci involved in the immune response are risk factors for hepatocellular carcinoma. Hepatology 52:2034-2043.

Cooper GM, Nickerson DA, Eichler EE. 2007. Mutational and selective effects on copy-number variants in the human genome. Nat Genet 39:S22-29.

Coquelle A, Toledo F, Stern S, Bieth A, Debatisse M. 1998. A new role for hypoxia in tumor progression: induction of fragile site triggering genomic rearrangements and formation of complex DMs and HSRs. Mol Cell 2:259-265.

Cordaux R, Batzer MA. 2009. The impact of retrotransposons on human genome evolution. Nat Rev Genet 10:691-703.

Das M, Garlick DS, Greiner DL, Davis RJ. 2011. The role of JNK in the development of hepatocellular carcinoma. Genes Dev 25:634-645.

Davis RJ. 2000. Signal transduction by the JNK group of MAP kinases. Cell 103:239-252.

Davit-Spraul A, Gonzales E, Baussan C, Jacquemin E. 2009. Progressive familial intrahepatic cholestasis. Orphanet J Rare Dis 4:1.

Davit-Spraul A, Gonzales E, Baussan C, Jacquemin E. 2010. The spectrum of liver diseases related to ABCB4 gene mutations: pathophysiology and clinical aspects. Semin Liver Dis 30:134-146.

de Cid R, Riveira-Munoz E, Zeeuwen PL, Robarge J, Liao W, Dannhauser EN, Giardina E, Stuart PE, Nair R, Helms C, et al. 2009. Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. Nat Genet 41:211-215.

Dereli-Oz A, Versini G, Halazonetis TD. 2011. Studies of genomic copy number changes in human cancers reveal signatures of DNA replication stress. Mol Oncol 5:308-314.

Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, et al. 2013. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform 14:671-683.

Diskin SJ, Hou C, Glessner JT, Attiyeh EF, Laudenslager M, Bosse K, Cole K, Mosse YP, Wood A, Lynch JE, et al. 2009. Copy number variation at 1q21.1 associated with neuroblastoma. Nature 459:987-991.

Downs JA, Nussenzweig MC, Nussenzweig A. 2007. Chromatin dynamics and the preservation of genetic information. Nature 447:951-958.

Dragani TA. 2010. Risk of HCC: genetic heterogeneity and complex genetics. J Hepatol 52:252-257.

Dvorak Z, Vrzal R, Henklova P, Jancova P, Anzenbacherova E, Maurel P, Svecova L, Pavek P, Ehrmann J, Havlik R, et al. 2008. JNK inhibitor SP600125 is a partial agonist of human aryl hydrocarbon receptor and induces CYP1A1 and CYP1A2 genes in primary human hepatocytes. Biochem Pharmacol 75:580-588.

Eferl R, Wagner EF. 2003. AP-1: a double-edged sword in tumorigenesis. Nat Rev Cancer 3:859-868.

El-Serag HB, Rudolph KL. 2007. Hepatocellular carcinoma: epidemiology and molecular carcinogenesis. Gastroenterology 132:2557-2576.

Elinav E, Nowarski R, Thaiss CA, Hu B, Jin C, Flavell RA. 2013. Inflammation-induced cancer: crosstalk between tumours, immune cells and microorganisms. Nat Rev Cancer 13:759-771.

Eltzschig HK, Carmeliet P. 2011. Hypoxia and inflammation. N Engl J Med 364:656-665.

Fanciulli M, Norsworthy PJ, Petretto E, Dong R, Harper L, Kamesh L, Heward JM, Gough SC, de Smith A, Blakemore AI, et al. 2007. FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. Nat Genet 39:721-723.

Farazi PA, DePinho RA. 2006. Hepatocellular carcinoma pathogenesis: from genes to environment. Nat Rev Cancer 6:674-687.

Fellermann K, Stange DE, Schaeffeler E, Schmalzl H, Wehkamp J, Bevins CL, Reinisch W, Teml A, Schwab M, Lichter P, et al. 2006. A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. Am J Hum Genet 79:439-448.

Ferlay J SI, Ervik M, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray, F. 2012. GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11[Internet]. Lyon, France: International Agency for Research on Cancer; 2013. Available from: http://globocan.iarc.fr, accessed on day/month/year.

Feuk L, Carson AR, Scherer SW. 2006. Structural variation in the human genome. Nat Rev Genet 7:85-97.

Fujimoto A, Totoki Y, Abe T, Boroevich KA, Hosoda F, Nguyen HH, Aoki M, Hosono N, Kubo M, Miya F, et al. 2012. Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. Nat Genet 44:760-764.

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. 2004. A census of human cancer genes. Nat Rev Cancer 4:177-183.

Geiger T, Cox J, Mann M. 2010. Proteomic changes resulting from gene copy number variations in cancer cells. PLoS Genet 6:e1001090.

Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, Nibbs RJ, Freedman BI, Quinones MP, Bamshad MJ, et al. 2005. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 307:1434-1440.

Gordon DJ, Resio B, Pellman D. 2012. Causes and consequences of aneuploidy in cancer. Nat Rev Genet 13:189-203.

Greenman CD, Bignell G, Butler A, Edkins S, Hinton J, Beare D, Swamy S, Santarius T, Chen L, Widaa S, et al. 2010. PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. Biostatistics 11:164-175.

Greshock J, Feng B, Nogueira C, Ivanova E, Perna I, Nathanson K, Protopopov A, Weber BL, Chin L. 2007. A comparison of DNA copy number profiling platforms. Cancer Res 67:10173-10180.

Grivennikov SI, Greten FR, Karin M. 2010. Immunity, inflammation, and cancer. Cell 140:883-899.

Gu W, Zhang F, Lupski JR. 2008. Mechanisms for human genomic rearrangements. Pathogenetics 1:4.

Guichard C, Amaddeo G, Imbeaud S, Ladeiro Y, Pelletier L, Maad IB, Calderaro J, Bioulac-Sage P, Letexier M, Degos F, et al. 2012. Integrated analysis of somatic mutations

and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. Nat Genet 44:694-698.

Gupta PK, Sahota A, Boyadjiev SA, Bye S, Shao C, O'Neill JP, Hunter TC, Albertini RJ, Stambrook PJ, Tischfield JA. 1997. High frequency in vivo loss of heterozygosity is primarily a consequence of mitotic recombination. Cancer Res 57:1188-1193.

Guryev V, Saar K, Adamovic T, Verheul M, van Heesch SA, Cook S, Pravenec M, Aitman T, Jacob H, Shull JD, et al. 2008. Distribution and functional impact of DNA copy number variation in the rat. Nat Genet 40:538-545.

Hagenbuch B, Meier PJ. 2004. Organic anion transporting polypeptides of the OATP/ SLC21 family: phylogenetic classification as OATP/ SLCO superfamily, new nomenclature and molecular/functional properties. Pflugers Arch 447:653-665.

Hamid AS, Tesfamariam IG, Zhang Y, Zhang ZG. 2013. Aflatoxin B1-induced hepatocellular carcinoma in developing countries: Geographical distribution, mechanism of action and prevention. Oncol Lett 5:1087-1092.

Hanahan D, Weinberg RA. 2000. The hallmarks of cancer. Cell 100:57-70.

Hanahan D, Weinberg RA. 2011. Hallmarks of cancer: the next generation. Cell 144:646-674.

Harismendy O, Frazer K. 2009. Method for improving sequence coverage uniformity of targeted genomic intervals amplified by LR-PCR using Illumina GA sequencing-by-synthesis technology. Biotechniques 46:229-231.

Hastings PJ, Ira G, Lupski JR. 2009. A microhomology-mediated break-induced replication model for the origin of human copy number variation. PLoS Genet 5:e1000327.

Hastings PJ, Lupski JR, Rosenberg SM, Ira G. 2009. Mechanisms of change in gene copy number. Nat Rev Genet 10:551-564.

He G, Yu GY, Temkin V, Ogata H, Kuntzen C, Sakurai T, Sieghart W, Peck-Radosavljevic M, Leffert HL, Karin M. 2010. Hepatocyte IKKbeta/NF-kappaB inhibits tumor promotion and progression by preventing oxidative stress-driven STAT3 activation. Cancer Cell 17:286-297.

Helbig I, Mefford HC, Sharp AJ, Guipponi M, Fichera M, Franke A, Muhle H, de Kovel C, Baker C, von Spiczak S, et al. 2009. 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy. Nat Genet 41:160-162.

Henrichsen CN, Chaignat E, Reymond A. 2009. Copy number variants, diseases and gene expression. Hum Mol Genet 18:R1-8.

Henrichsen CN, Vinckenbosch N, Zollner S, Chaignat E, Pradervand S, Schutz F, Ruedi M, Kaessmann H, Reymond A. 2009. Segmental copy number variation shapes tissue transcriptomes. Nat Genet 41:424-429.

Higgs DR, Vickers MA, Wilkie AO, Pretorius IM, Jarman AP, Weatherall DJ. 1989. A review of the molecular genetics of the human alpha-globin gene cluster. Blood 73:1081-1104.

Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ, et al. 2007. Genome-wide in situ exon capture for selective resequencing. Nat Genet 39:1522-1527.

Huang J, Deng Q, Wang Q, Li KY, Dai JH, Li N, Zhu ZD, Zhou B, Liu XY, Liu RF, et al. 2012. Exome sequencing of hepatitis B virus-associated hepatocellular carcinoma. Nat Genet 44:1117-1121.

Hui L, Bakiri L, Mairhorfer A, Schweifer N, Haslinger C, Kenner L, Komnenovic V, Scheuch H, Beug H, Wagner EF. 2007. p38alpha suppresses normal and cancer cell proliferation by antagonizing the JNK-c-Jun pathway. Nat Genet 39:741-749.

Hui L, Zatloukal K, Scheuch H, Stepniak E, Wagner EF. 2008. Proliferation of human HCC cells and chemically induced mouse liver cancers requires JNK1-dependent p21 downregulation. J Clin Invest 118:3943-3953.

Hussain SP, Schwank J, Staib F, Wang XW, Harris CC. 2007. TP53 mutations and hepatocellular carcinoma: insights into the etiology and pathogenesis of liver cancer. Oncogene 26:2166-2176.

International Schizophrenia C. 2008. Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature 455:237-241.

Ioannidis JP, Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, Falchi M, Furlanello C, Game L, Jurman G, et al. 2009. Repeatability of published microarray gene expression analyses. Nat. Genet. 41:149-155.

Jaco I, Canela A, Vera E, Blasco MA. 2008. Centromere mitotic recombination in mammalian cells. J Cell Biol 181:885-892.

Jacquemin E. 2012. Progressive familial intrahepatic cholestasis. Clin Res Hepatol Gastroenterol 36 Suppl 1:S26-35.

Jang JJ, Weghorst CM, Henneman JR, Devor DE, Ward JM. 1992. Progressive atypia in spontaneous and N-nitrosodiethylamine-induced hepatocellular adenomas of C3H/HeNCr mice. Carcinogenesis 13:1541-1547.

Jensen LJ, Julien P, Kuhn M, von Mering C, Muller J, Doerks T, Bork P. 2008. eggNOG: automated construction and annotation of orthologous groups of genes. Nucleic Acids Res 36:D250-254.

Jiang Z, Jhunjhunwala S, Liu J, Haverty PM, Kennemer MI, Guan Y, Lee W, Carnevali P, Stinson J, Johnson S, et al. 2012. The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients. Genome Res 22:593-601.

Kabbarah O, Nogueira C, Feng B, Nazarian RM, Bosenberg M, Wu M, Scott KL, Kwong LN, Xiao Y, Cordon-Cardo C, et al. 2010. Integrative genome comparison of primary and metastatic melanomas. PLoS One 5:e10770.

Kamburov A, Stelzl U, Lehrach H, Herwig R. 2013. The ConsensusPathDB interaction database: 2013 update. Nucleic Acids Res 41:D793-800.

Kartenbeck J, Leuschner U, Mayer R, Keppler D. 1996. Absence of the canalicular isoform of the MRP gene-encoded conjugate export pump from the hepatocytes in Dubin-Johnson syndrome. Hepatology 23:1061-1066.

Katzenellenbogen M, Mizrahi L, Pappo O, Klopstock N, Olam D, Jacob-Hirsch J, Amariglio N, Rechavi G, Domany E, Galun E, et al. 2007. Molecular mechanisms of liver carcinogenesis in the mdr2-knockout mice. Mol Cancer Res 5:1159-1170.

Katzenellenbogen M, Pappo O, Barash H, Klopstock N, Mizrahi L, Olam D, Jacob-Hirsch J, Amariglio N, Rechavi G, Mitchell LA, et al. 2006. Multiple adaptive mechanisms to chronic liver disease revealed at early stages of liver carcinogenesis in the Mdr2-knockout mice. Cancer Res 66:4001-4010.

Keitel V, Kartenbeck J, Nies AT, Spring H, Brom M, Keppler D. 2000. Impaired protein maturation of the conjugate export pump multidrug resistance protein 2 as a consequence of a deletion mutation in Dubin-Johnson syndrome. Hepatology 32:1317-1328.

Kim TM, Xi R, Luquette LJ, Park RW, Johnson MD, Park PJ. 2013. Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes. Genome Res 23:217-227.

Knierim E, Lucke B, Schwarz JM, Schuelke M, Seelow D. 2011. Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. PLoS One 6:e28240.

Knisely AS, Strautnieks SS, Meier Y, Stieger B, Byrne JA, Portmann BC, Bull LN, Pawlikowska L, Bilezikci B, Ozcay F, et al. 2006. Hepatocellular carcinoma in ten children under five years of age with bile salt export pump deficiency. Hepatology 44:478-486.

Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 22:568-576.

Korbel JO, Abyzov A, Mu XJ, Carriero N, Cayting P, Zhang Z, Snyder M, Gerstein MB. 2009. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. Genome Biol 10:R23.

Korbel JO, Campbell PJ. 2013. Criteria for inference of chromothripsis in cancer genomes. Cell 152:1226-1236.

Kremsdorf D, Soussan P, Paterlini-Brechot P, Brechot C. 2006. Hepatitis B virus-related hepatocellular carcinoma: paradigms for viral-related human carcinogenesis. Oncogene 25:3823-3833.

Krepischi AC, Achatz MI, Santos EM, Costa SS, Lisboa BC, Brentani H, Santos TM, Goncalves A, Nobrega AF, Pearson PL, et al. 2012. Germline DNA copy number variation in familial and early-onset breast cancer. Breast Cancer Res 14:R24.

Krepischi AC, Pearson PL, Rosenberg C. 2012. Germline copy number variations and cancer predisposition. Future Oncol 8:441-450.

Kumareswaran R, Ludkovski O, Meng A, Sykes J, Pintilie M, Bristow RG. 2012. Chronic hypoxia compromises repair of DNA double-strand breaks to drive genetic instability. J Cell Sci 125:189-199.

Lan X, Rai P, Chandel N, Cheng K, Lederman R, Saleem MA, Mathieson PW, Husain M, Crosson JT, Gupta K, et al. 2013. Morphine induces albuminuria by compromising podocyte integrity. PLoS One 8:e55748.

Laurent-Puig P, Zucman-Rossi J. 2006. Genetics of hepatocellular tumors. Oncogene 25:3778-3786.

Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, et al. 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499:214-218.

Le Marechal C, Masson E, Chen JM, Morel F, Ruszniewski P, Levy P, Ferec C. 2006. Hereditary pancreatitis caused by triplication of the trypsinogen locus. Nat Genet 38:1372-1374.

Lee JA, Carvalho CM, Lupski JR. 2007. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell 131:1235-1247.

Levrero M. 2006. Viral hepatitis and liver cancer: the case of hepatitis C. Oncogene 25:3834-3847.

Li A, Liu Z, Lezon-Geyda K, Sarkar S, Lannin D, Schulz V, Krop I, Winer E, Harris L, Tuck D. 2011. GPHMM: an integrated hidden Markov model for identification of copy number alteration and loss of heterozygosity in complex tumor samples using whole genome SNP arrays. Nucleic Acids Res 39:4928-4941.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-2079.

Li M, Zhao H, Zhang X, Wood LD, Anders RA, Choti MA, Pawlik TM, Daniel HD, Kannangai R, Offerhaus GJ, et al. 2011. Inactivating mutations of the chromatin remodeling gene ARID2 in hepatocellular carcinoma. Nat Genet 43:828-829.

Li QL, Ito K, Sakakura C, Fukamachi H, Inoue K, Chi XZ, Lee KY, Nomura S, Lee CW, Han SB, et al. 2002. Causal relationship between the loss of RUNX3 expression and gastric cancer. Cell 109:113-124.

Lieber MR. 2008. The mechanism of human nonhomologous DNA end joining. J Biol Chem 283:1-5.

Liu B, Morrison CD, Johnson CS, Trump DL, Qin M, Conroy JC, Wang J, Liu S. 2013. Computational methods for detecting copy number variations in cancer genome using next generation sequencing: principles and challenges. Oncotarget 4:1868-1881.

Liu B, Yang L, Huang B, Cheng M, Wang H, Li Y, Huang D, Zheng J, Li Q, Zhang X, et al. 2012. A functional copy-number variation in MAPKAPK2 predicts risk and prognosis of lung cancer. Am J Hum Genet 91:384-390.

Liu W, Sun J, Li G, Zhu Y, Zhang S, Kim ST, Sun J, Wiklund F, Wiley K, Isaacs SD, et al. 2009. Association of a germ-line copy number variation at 2p24.3 and risk for aggressive prostate cancer. Cancer Res 69:2176-2179.

Llovet JM. 2005. Updated treatment approach to hepatocellular carcinoma. J Gastroenterol 40:225-235.

Lu K, Lee MH, Hazard S, Brooks-Wilson A, Hidaka H, Kojima H, Ose L, Stalenhoef AF, Mietinnen T, Bjorkhem I, et al. 2001. Two genes that map to the STSL locus cause sitosterolemia: genomic structure and spectrum of mutations involving sterolin-1 and sterolin-2, encoded by ABCG5 and ABCG8, respectively. Am J Hum Genet 69:278-290.

Lucito R, Suresh S, Walter K, Pandey A, Lakshmi B, Krasnitz A, Sebat J, Wigler M, Klein AP, Brune K, et al. 2007. Copy-number variants in patients with a strong family history of pancreatic cancer. Cancer Biol Ther 6:1592-1599.

Lupski JR. 2004. Hotspots of homologous recombination in the human genome: not all homologous sequences are equal. Genome Biol 5:242.

Maeda S, Kamata H, Luo JL, Leffert H, Karin M. 2005. IKKbeta couples hepatocyte death to cytokine-driven compensatory proliferation that promotes chemical hepatocarcinogenesis. Cell 121:977-990.

Magi A, Tattini L, Cifola I, D'Aurizio R, Benelli M, Mangano E, Battaglia C, Bonora E, Kurg A, Seri M, et al. 2013. EXCAVATOR: detecting copy number variants from whole-exome sequencing data. Genome Biol 14:R120.

Magi A, Tattini L, Pippucci T, Torricelli F, Benelli M. 2012. Read count approach for DNA copy number variants detection. Bioinformatics 28:470-478.

Malhotra D, Sebat J. 2012. CNVs: harbingers of a rare variant revolution in psychiatric genetics. Cell 148:1223-1241.

Manning BD, Cantley LC. 2007. AKT/PKB signalling: navigating downstream. Cell 129:1261-1274.

Mantovani A, Allavena P, Sica A, Balkwill F. 2008. Cancer-related inflammation. Nature 454:436-444.

Marozin S, Altomonte J, Apfel S, Dinh PX, De Toni EN, Rizzani A, Nussler A, Kato N, Schmid RM, Pattnaik AK, et al. 2012. Posttranslational modification of vesicular stomatitis virus glycoprotein, but not JNK inhibition, is the antiviral mechanism of SP600125. J Virol 86:4844-4855.

Marra F, Gastaldelli A, Svegliati Baroni G, Tell G, Tiribelli C. 2008. Molecular basis and mechanisms of progression of non-alcoholic steatohepatitis. Trends Mol Med 14:72-81.

Mauad TH, van Nieuwkerk CM, Dingemans KP, Smit JJ, Schinkel AH, Notenboom RG, van den Bergh Weerman MA, Verkruisen RP, Groen AK, Oude Elferink RP, et al. 1994. Mice with homozygous disruption of the mdr2 P-glycoprotein gene. A novel animal model for studies of nonsuppurative inflammatory cholangitis and hepatocarcinogenesis. Am J Pathol 145:1237-1245.

Maurici D, Perez-Atayde A, Grier HE, Baldini N, Serra M, Fletcher JA. 1998. Frequency and implications of chromosome 8 and 12 gains in Ewing sarcoma. Cancer Genet Cytogenet 100:106-110.

McCarroll SA, Huett A, Kuballa P, Chilewski SD, Landry A, Goyette P, Zody MC, Hall JL, Brant SR, Cho JH, et al. 2008. Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. Nat Genet 40:1107-1112.

McGivern DR, Lemon SM. 2011. Virus-specific mechanisms of carcinogenesis in hepatitis C virus associated liver cancer. Oncogene 30:1969-1983.

McVey M, Lee SE. 2008. MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. Trends Genet 24:529-538.

Mefford HC, Sharp AJ, Baker C, Itsara A, Jiang Z, Buysse K, Huang S, Maloney VK, Crolla JA, Baralle D, et al. 2008. Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. N Engl J Med 359:1685-1699.

Naugler WE, Karin M. 2008. The wolf in sheep's clothing: the role of interleukin-6 in immunity, inflammation and cancer. Trends Mol Med 14:109-119.

Nguyen DQ, Webber C, Ponting CP. 2006. Bias of selection on human copy-number variants. PLoS Genet 2:e20.

Nikolaou K, Tsagaratou A, Eftychi C, Kollias G, Mosialos G, Talianidis I. 2012. Inactivation of the deubiquitinase CYLD in hepatocytes causes apoptosis, inflammation, fibrosis, and cancer. Cancer Cell 21:738-750.

Ohno S. 1970. Evolution by gene duplication. Berlin-Heidelberg-New York, Springer-Verlang.

Pauli-Magnus C, Kerb R, Fattinger K, Lang T, Anwald B, Kullak-Ublick GA, Beuers U, Meier PJ. 2004. BSEP and MDR3 haplotype structure in healthy Caucasians, primary biliary cirrhosis and primary sclerosing cholangitis. Hepatology 39:779-791.

Pauli-Magnus C, Meier PJ. 2006. Hepatobiliary transporters and drug-induced cholestasis. Hepatology 44:778-787.

Pauli-Magnus C, Meier PJ. 2005. Hepatocellular transporters and cholestasis. J Clin Gastroenterol 39:S103-110.

Paulsson K, Johansson B. 2007. Trisomy 8 as the sole chromosomal aberration in acute myeloid leukemia and myelodysplastic syndromes. Pathol Biol (Paris) 55:37-48.

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, et al. 2007. Diet and the evolution of human amylase gene copy number variation. Nat Genet 39:1256-1260.

Perry GH, Yang F, Marques-Bonet T, Murphy C, Fitzgerald T, Lee AS, Hyland C, Stone AC, Hurles ME, Tyler-Smith C, et al. 2008. Copy number variation and evolution in humans and chimpanzees. Genome Res 18:1698-1710.

Perz JF, Armstrong GL, Farrington LA, Hutin YJ, Bell BP. 2006. The contributions of hepatitis B virus and hepatitis C virus infections to cirrhosis and primary liver cancer worldwide. J Hepatol 45:529-538.

Pikarsky E, Porat RM, Stein I, Abramovitch R, Amit S, Kasem S, Gutkovich-Pyest E, Urieli-Shoval S, Galun E, Ben-Neriah Y. 2004. NF-kappaB functions as a tumour promoter in inflammation-associated cancer. Nature 431:461-466.

Popova T, Manie E, Stoppa-Lyonnet D, Rigaill G, Barillot E, Stern MH. 2009. Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays. Genome Biol 10:R128.

Qi H, Dal Cin P, Hernandez JM, Garcia JL, Sciot R, Fletcher C, Van Eyken P, De Wever I, Van den Berghe H. 1996. Trisomies 8 and 20 in desmoid tumors. Cancer Genet Cytogenet 92:147-149.

Quackenbush J. 2002. Microarray data normalization and transformation. Nat Genet 32 Suppl:496-501.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841-842.

Raeymaekers P, Timmerman V, Nelis E, De Jonghe P, Hoogendijk JE, Baas F, Barker DF, Martin JJ, De Visser M, Bolhuis PA, et al. 1991. Duplication in chromosome 17p11.2 in Charcot-Marie-Tooth neuropathy type 1a (CMT 1a). The HMSN Collaborative Research Group. Neuromuscul Disord 1:93-97.

Raman M, Chen W, Cobb MH. 2007. Differential regulation and properties of MAPKs. Oncogene 26:3100-3112.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al. 2006. Global variation in copy number in the human genome. Nature 444:444-454.

Rovelet-Lecrux A, Hannequin D, Raux G, Le Meur N, Laquerriere A, Vital A, Dumanchin C, Feuillette S, Brice A, Vercelletto M, et al. 2006. APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. Nat Genet 38:24-26.

Sakurai T, Maeda S, Chang L, Karin M. 2006. Loss of hepatic NF-kappa B activity enhances chemical hepatocarcinogenesis through sustained c-Jun N-terminal kinase 1 activation. Proc Natl Acad Sci U S A 103:10544-10551.

Sathirapongsasuti JF, Lee H, Horst BA, Brunner G, Cochran AJ, Binder S, Quackenbush J, Nelson SF. 2011. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. Bioinformatics 27:2648-2654.

Scheimann AO, Strautnieks SS, Knisely AS, Byrne JA, Thompson RJ, Finegold MJ. 2007. Mutations in bile salt export pump (ABCB11) in two children with progressive familial intrahepatic cholestasis and cholangiocarcinoma. J Pediatr 150:556-559.

Schinkel AH, Wagenaar E, van Deemter L, Mol CA, Borst P. 1995. Absence of the mdr1a P-Glycoprotein in mice affects tissue distribution and pharmacokinetics of dexamethasone, digoxin, and cyclosporin A. J Clin Invest 96:1698-1705.

Seeger RC, Brodeur GM, Sather H, Dalton A, Siegel SE, Wong KY, Hammond D. 1985. Association of multiple copies of the N-myc oncogene with rapid progression of neuroblastomas. N Engl J Med 313:1111-1116.

Seitz HK, Stickel F. 2007. Molecular mechanisms of alcohol-mediated carcinogenesis. Nat Rev Cancer 7:599-612.

Sharp AJ, Mefford HC, Li K, Baker C, Skinner C, Stevenson RE, Schroer RJ, Novara F, De Gregori M, Ciccone R, et al. 2008. A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. Nat Genet 40:322-328.

Shibata S, Tada Y, Asano Y, Hau CS, Kato T, Saeki H, Yamauchi T, Kubota N, Kadowaki T, Sato S. 2012. Adiponectin regulates cutaneous wound healing by promoting keratinocyte proliferation and migration via the ERK signalling pathway. J Immunol 189:3231-3241.

Shlien A, Malkin D. 2010. Copy number variations and cancer susceptibility. Curr Opin Oncol 22:55-63.

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. 2014. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet 15:121-132.

Singleton AB, Farrer M, Johnson J, Singleton A, Hague S, Kachergus J, Hulihan M, Peuralinna T, Dutra A, Nussbaum R, et al. 2003. alpha-Synuclein locus triplication causes Parkinson's disease. Science 302:841.

Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL. 1987. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. Science 235:177-182.

Smit JJ, Schinkel AH, Oude Elferink RP, Groen AK, Wagenaar E, van Deemter L, Mol CA, Ottenhoff R, van der Lugt NM, van Roon MA, et al. 1993. Homozygous disruption of the murine mdr2 P-glycoprotein gene leads to a complete absence of phospholipid from bile and to liver disease. Cell 75:451-462.

Stankiewicz P, Lupski JR. 2002. Genome architecture, rearrangements and genomic disorders. Trends Genet 18:74-82.

Stefansson H, Rujescu D, Cichon S, Pietilainen OP, Ingason A, Steinberg S, Fossdal R, Sigurdsson E, Sigmundsson T, Buizer-Voskamp JE, et al. 2008. Large recurrent microdeletions associated with schizophrenia. Nature 455:232-236.

Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA, et al. 2011. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell 144:27-40.

Strautnieks SS, Byrne JA, Pawlikowska L, Cebecauerova D, Rayner A, Dutton L, Meier Y, Antoniou A, Stieger B, Arnell H, et al. 2008. Severe bile salt export pump deficiency: 82 different ABCB11 mutations in 109 families. Gastroenterology 134:1203-1214.

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A 101:6062-6067.

Sung WK, Zheng H, Li S, Chen R, Liu X, Li Y, Lee NP, Lee WH, Ariyaratne PN, Tennakoon C, et al. 2012. Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. Nat Genet 44:765-769.

Tan Z, Qian X, Jiang R, Liu Q, Wang Y, Chen C, Wang X, Ryffel B, Sun B. 2013. IL-17A plays a critical role in the pathogenesis of liver fibrosis through hepatic stellate cell activation. J Immunol 191:1835-1844.

Tang YC, Amon A. 2013. Gene copy-number alterations: a cost-benefit analysis. Cell 152:394-405.

Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, Varambally S, Cao X, Tchinda J, Kuefer R, et al. 2005. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science 310:644-648.

Tonon G, Wong KK, Maulik G, Brennan C, Feng B, Zhang Y, Khatry DB, Protopopov A, You MJ, Aguirre AJ, et al. 2005. High-resolution genomic profiles of human lung cancer. Proc Natl Acad Sci U S A 102:9625-9630.

Totoki Y, Tatsuno K, Yamamoto S, Arai Y, Hosoda F, Ishikawa S, Tsutsumi S, Sonoda K, Totsuka H, Shirakihara T, et al. 2011. High-resolution characterization of a hepatocellular carcinoma genome. Nat Genet 43:464-469.

Tournier C, Dong C, Turner TK, Jones SN, Flavell RA, Davis RJ. 2001. MKK7 is an essential component of the JNK signal transduction pathway activated by proinflammatory cytokines. Genes Dev 15:1419-1426.

Trotman LC, Niki M, Dotan ZA, Koutcher JA, Di Cristofano A, Xiao A, Khoo AS, Roy-Burman P, Greenberg NM, Van Dyke T, et al. 2003. Pten dose dictates cancer progression in the prostate. PLoS Biol 1:E59.

Tse KP, Su WH, Yang ML, Cheng HY, Tsang NM, Chang KP, Hao SP, Yao Shugart Y, Chang YS. 2011. A gender-specific association of CNV at 6p21.3 with NPC susceptibility. Hum Mol Genet 20:2889-2896.

Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, et al. 2005. Fine-scale structural variation of the human genome. Nat Genet 37:727-732.

Ueda H, Ullrich SJ, Gangemi JD, Kappel CA, Ngo L, Feitelson MA, Jay G. 1995. Functional inactivation but not structural mutation of p53 causes liver cancer. Nat Genet 9:41-47.

Ulrich CM, Bigler J, Potter JD. 2006. Non-steroidal anti-inflammatory drugs for cancer prevention: promise, perils and pharmacogenetics. Nat Rev Cancer 6:130-140.

Unsal H, Yakicier C, Marcais C, Kew M, Volkmann M, Zentgraf H, Isselbacher KJ, Ozturk M. 1994. Genetic heterogeneity of hepatocellular carcinoma. Proc Natl Acad Sci U S A 91:822-826.

Van Loo P, Nordgard SH, Lingjaerde OC, Russnes HG, Rye IH, Sun W, Weigman VJ, Marynen P, Zetterberg A, Naume B, et al. 2010. Allele-specific copy number analysis of tumors. Proc Natl Acad Sci U S A 107:16910-16915.

van Mil SW, van der Woerd WL, van der Brugge G, Sturm E, Jansen PL, Bull LN, van den Berg IE, Berger R, Houwen RH, Klomp LW. 2004. Benign recurrent intrahepatic cholestasis type 2 is caused by mutations in ABCB11. Gastroenterology 127:379-384.

Venkatachalam R, Verwiel ET, Kamping EJ, Hoenselaar E, Gorgens H, Schackert HK, van Krieken JH, Ligtenberg MJ, Hoogerbrugge N, van Kessel AG, et al. 2011. Identification of candidate predisposing copy number variants in familial and early-onset colorectal cancer patients. Int J Cancer 129:1635-1642.

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Jr., Kinzler KW. 2013. Cancer genome landscapes. Science 339:1546-1558.

Volker M, Backstrom N, Skinner BM, Langley EJ, Bunzey SK, Ellegren H, Griffin DK. 2010. Copy number variation, chromosome rearrangement, and their association with recombination during avian evolution. Genome Res 20:503-511.

Wada T, Joza N, Cheng HY, Sasaki T, Kozieradzki I, Bachmaier K, Katada T, Schreiber M, Wagner EF, Nishina H, et al. 2004. MKK7 couples stress signalling to G2/M cell-cycle progression and cellular senescence. Nat Cell Biol 6:215-226.

Walsh KM, Choi M, Oberg K, Kulke MH, Yao JC, Wu C, Jurkiewicz M, Hsu LI, Hooshmand SM, Hassan M, et al. 2011. A pilot genome-wide association study shows genomic variants enriched in the non-tumor cells of patients with well-differentiated neuroendocrine tumors of the ileum. Endocr Relat Cancer 18:171-180.

Weischenfeldt J, Symmons O, Spitz F, Korbel JO. 2013. Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet 14:125-138.

Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, Fossdal R, Saemundsen E, Stefansson H, Ferreira MA, Green T, et al. 2008. Association between microdeletion and microduplication at 16p11.2 and autism. N Engl J Med 358:667-675.

Weston CR, Davis RJ. 2007. The JNK signal transduction pathway. Curr Opin Cell Biol 19:142-149.

Willer CJ, Speliotes EK, Loos RJ, Li S, Lindgren CM, Heid IM, Berndt SI, Elliott AL, Jackson AU, Lamina C, et al. 2009. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nat Genet 41:25-34.

Wolfe KH. 2001. Yesterday's polyploids and the mystery of diploidization. Nat Rev Genet 2:333-341.

Yang XR, Brown K, Landi MT, Ghiorzo P, Badenas C, Xu M, Hayward NK, Calista D, Landi G, Bruno W, et al. 2012. Duplication of CXC chemokine genes on chromosome 4q13 in a melanoma-prone family. Pigment Cell Melanoma Res 25:243-247.

Yang Y, Chung EK, Wu YL, Savelli SL, Nagaraja HN, Zhou B, Hebert M, Jones KN, Shu Y, Kitzmiller K, et al. 2007. Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. Am J Hum Genet 80:1037-1054.

Yau C, Mouradov D, Jorissen RN, Colella S, Mirza G, Steers G, Harris A, Ragoussis J, Sieber O, Holmes CC. 2010. A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. Genome Biol 11:R92.

Ye L, Kleiner S, Wu J, Sah R, Gupta RK, Banks AS, Cohen P, Khandekar MJ, Bostrom P, Mepani RJ, et al. 2012. TRPV4 is a regulator of adipose oxidative metabolism, inflammation, and energy homeostasis. Cell 151:96-110.

Yea S, Narla G, Zhao X, Garg R, Tal-Kremer S, Hod E, Villanueva A, Loke J, Tarocchi M, Akita K, et al. 2008. Ras promotes growth by alternative splicing-mediated inactivation of the KLF6 tumor suppressor in hepatocellular carcinoma. Gastroenterology 134:1521-1531.

Yoshida K, Sanada M, Shiraishi Y, Nowak D, Nagata Y, Yamamoto R, Sato Y, Sato-Otsubo A, Kon A, Nagasaki M, et al. 2011. Frequent pathway mutations of splicing machinery in myelodysplasia. Nature 478:64-69.

Yoshimoto M, Cutz JC, Nuin PA, Joshua AM, Bayani J, Evans AJ, Zielenska M, Squire JA. 2006. Interphase FISH analysis of PTEN in histologic sections shows genomic

deletions in 68% of primary prostate cancer and 23% of high-grade prostatic intra-epithelial neoplasias. Cancer Genet Cytogenet 169:128-137.

Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhang CZ, Wala J, Mermel CH, et al. 2013. Pan-cancer patterns of somatic copy number alteration. Nat Genet 45:1134-1140.

Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, Chambers MC, Zimmerman LJ, Shaddox KF, Kim S, et al. 2014. Proteogenomic characterization of human colon and rectal cancer. Nature 513:382-387.

Zhang C, Zhang C, Chen S, Yin X, Pan X, Lin G, Tan Y, Tan K, Xu Z, Hu P, et al. 2013. A single cell level based method for copy number variation analysis by low coverage massively parallel sequencing. PLoS One 8:e54236.

Zhang F, Gu W, Hurles ME, Lupski JR. 2009. Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet 10:451-481.

Zhang Z. 2012. Genomic landscape of liver cancer. Nat Genet 44:1075-1077.

Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. 2013. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics 14 Suppl 11:S1.

Zheng L, Lee WH. 2002. Retinoblastoma tumor suppressor and genome stability. Adv Cancer Res 85:13-50.

Zhuang Z, Park WS, Pack S, Schmidt L, Vortmeyer AO, Pak E, Pham T, Weil RJ, Candidus S, Lubensky IA, et al. 1998. Trisomy 7-harbouring non-random duplication of the mutant MET allele in hereditary papillary renal carcinomas. Nat Genet 20:66-69.

Acknowledgements

This thesis would not have been possible without the help and support of the many kind people, to only some of whom it is possible to give particular mention here.

I would first and foremost like to express my sincere gratitude to Francesca Ciccarelli for the continuous support and encouragement throughout these four years of PhD. Her guidance had helped me become a better person and hopefully a good scientist.

I would like to thank my external co-supervisor Tomás Marquès-Bonet for taking time to review my reports and give critical comments and suggestions.

I would like to thank my internal co-supervisor Gioacchino Natoli for his support and useful discussions during the course of the PhD.

I would like to acknowledge the financial, academic and technical support of MODHEP consortium, Lifelong Learning Programme-Erasmus Placement of the University of Milan, European School of Molecular Medicine and Istituto Europeo di Oncologia. Particularly, the staff members for their help with the administrative work.

I am very grateful to all the members of the group, Anna De Grassi, Fabio Iannelli, Matteo Cereda, Gennaro Gambardella, Matteo D'Antonio, Vera Pendino, Elena Gatti, Omer An and Valentina Melocchi for their useful discussions on science and philosophy of life. Their support and pranks always made for good times and helped to sail through the stressful periods. I would like to make special mention of Fabio Iannelli for helping me with the biological aspects of the project.

Last, but by no means least, I thank my family and friends in India, Milano and elsewhere for their support and encouragement throughout. I would like to especially thank my mother, sister, brother and Sudharshan Elangovan, for helping me maintain my sanity and reminding me that there is more to life. A special mention of few of my friends, Archana Varadaraj, Janaina Oishi and Sriganesh Jammula for the shared drinks and fun times.

nature COMMUNICATIONS

ARTICLE

Received 13 Dec 2013 | Accepted 10 Apr 2014 | Published 13 May 2014

DOI: 10.1038/ncomms4850

Massive gene amplification drives paediatric hepatocellular carcinoma caused by bile salt export pump deficiency

Fabio Iannelli[1,*], Agnese Collino[1,*], Shruti Sinha[1,2,*], Enrico Radaelli[3], Paola Nicoli[1], Lorenzo D'Antiga[4], Aurelio Sonzogni[5], Jamila Faivre[6], Marie Annick Buendia[6], Ekkehard Sturm[7], Richard J. Thompson[8], A.S. Knisely[9], Gioacchino Natoli[1], Serena Ghisletti[1] & Francesca D. Ciccarelli[1,2]

Hepatocellular carcinoma (HCC) is almost invariably associated with an underlying inflammatory state, whose direct contribution to the acquisition of critical genomic changes is unclear. Here we map acquired genomic alterations in human and mouse HCCs induced by defects in hepatocyte biliary transporters, which expose hepatocytes to bile salts and cause chronic inflammation that develops into cancer. In both human and mouse cancer genomes, we find few somatic point mutations with no impairment of cancer genes, but massive gene amplification and rearrangements. This genomic landscape differs from that of virus- and alcohol-associated liver cancer. Copy-number gains preferentially occur at late stages of cancer development and frequently target the MAPK signalling pathway, and in particular direct regulators of JNK. The pharmacological inhibition of JNK retards cancer progression in the mouse. Our study demonstrates that intrahepatic cholestasis leading to hepatocyte exposure to bile acids and inflammation promotes cancer through genomic modifications that can be distinguished from those determined by other aetiological factors.

[1] European Institute of Oncology (IEO), Department of Experimental Oncology, IFOM-IEO Campus, Via Adamello 16, 20139 Milan, Italy. [2] Division of Cancer Studies, King's College London, London SE1 1UL, UK. [3] VIB Center for the Biology of Disease, KU Leuven Center for Human Genetics, O&N4 Herestraat 49 box 602, B-3000 Leuven, Belgium. [4] Paediatric Hepatology, Gastroenterology and Transplantation, Ospedale Papa Giovanni XXIII, Piazza OMS - Organizzazione Mondiale della Sanità 1, 24128 Bergamo, Italy. [5] Department of Pathology, Ospedale Papa Giovanni XXIII, Piazza OMS - Organizzazione Mondiale della Sanità 1, 24128 Bergamo, Italy. [6] Institut National de la Santé et de la Recherche Médicale (INSERM) U785, University Paris-Sud, France, Centre Hépatobiliaire, Hôpital Paul Brousse, Villejuif F94800, France. [7] University Hospital for Children and Adolescents, University of Tuebingen, 72076 Tuebingen, Germany. [8] Institute of Liver Studies, King's College London, London SE5 9RS, UK. [9] Institute of Liver Studies, King's College Hospital, London SE5 9RS, UK. * These authors contributed equally to this work. Correspondence and requests for materials should be addressed to F.D.C. (email: francesca.ciccarelli@kcl.ac.uk) or to S.G. (email: serena.ghisletti@ieo.eu).

Human hepatocellular carcinoma (HCC) arises in response to identifiable causes that in most patients show little or no overlap, including hepatitis C virus (HCV) or hepatitis B virus (HBV) infection, alcohol, exposure to aflatoxin and non-alcoholic steato-hepatitis associated with metabolic diseases[1,2]. Because of these distinct aetiologies, the characterization of different types of HCCs allows a basic question in cancer biology to be addressed, namely the specificity of the tumour genomic landscape relative to the disease causative factors. Recent data already suggested that the high genetic heterogeneity of liver cancer depends on the initiating agents. For example, genomic re-sequencing of human liver cancers highlighted recurrent mutations in key cancer genes, such as *TP53* and *CTNNB1*, and chromatin regulators, but also substantial differences[3–8]. Genes encoding components of the chromatin-remodelling complexes are frequently mutated in hepatitis HCV- but not in HBV-associated HCC[5]. Also the tumour mutational signature (that is, number and type of acquired mutations) strongly depends on the underlying mutagenic mechanism and exposure to different genotoxic chemicals leads to distinct mutation patterns[3,4,6]. In addition, integration of HBV DNA into the host genome induces genomic instability but may also directly modify cancer driver genes[6,8,9], thus triggering oncogenic events. These data clearly show that HCC has a complex pathogenesis in which exogenous factors, inflammation and sustained regeneration cooperate to promote cancer.

To understand how liver injury due to a combination of chemical damage, inflammation and fibrosis contributes to the acquired cancer genomic instability, we profiled the genome of human HCCs associated with bile salt export pump (BSEP) deficiency, also known as progressive familial intrahepatic cholestasis type 2. Progressive familial intrahepatic cholestasis designates a heterogeneous group of rare autosomal recessive disorders caused by inherited inactivating mutations in the hepatocyte membrane transporter genes *ATP8B1*, *ABCB11* and *ABCB4* (ref. 10). The disease usually appears in infancy or early childhood and manifests with hepatocellular damage and cholestasis due to defects in bile formation[10]. In BSEP deficiency, inherited mutations in the *ABCB11* gene cause impairment of bile salt export from hepatocytes into bile, leading to liver chronic inflammation and to the early onset of hepatocellular carcinoma[11]. Thus, this liver cancer type, which we refer to as BSEP-HCC, provides the opportunity to map the acquired genomic modifications that trigger liver cancer in the absence of external mutagenic factors.

The genomic profiling of BSEP-HCCs revealed a scenario different from all other HCCs sequenced to date. Exome sequencing showed only very few somatic mutations that did not affect known cancer genes. In contrast, BSEP-HCC genomes acquired massive gene amplification that affected components of signal transduction pathways, such as the ErbB, the PI3K/Akt and the mitogen-activated protein kinase (MAPK) signalling pathways. To further examine the role of these aberrations in cancer onset and progression, we re-sequenced exomes and genomes of HCCs from *Mdr2*-KO mice. The *Mdr2*-encoded P-glycoprotein belongs to the ABC family of membrane transporters and its absence impairs the secretion of phosphatidylcholine into biliary canaliculi[12]. The resulting high concentration of monomeric bile salts induces hepatocellular damage, inflammation and eventually HCC with high penetrance[13–16]. It has been suggested that this sequence of events recapitulates to some extent the development of the most common types of human HCCs. In fact, HCCs arisen in chronic liver disease in *Mdr2*-KO mice have an aetiopathogenesis similar to that of BSEP-HCCs, while it is clearly distinct from that of viral and metabolic disease-associated human HCCs. Consistent with this notion, in *Mdr2*-KO HCC we identified very few somatic point mutations, instead we detected a progressive accumulation of gene amplifications affecting the MAPK signalling pathways, and in particular activators of the cJun-N terminal kinases (JNK). Pharmacological inhibition of JNK in the mouse dampened cancer progression.

Altogether, our data suggest that the genetic heterogeneity of liver cancers correlates with their distinct pathogenesis and leads to different mechanisms of tumour progression that can be at least partially recapitulated in the mouse. They also provide evidence that solid tumours are not necessarily associated with mutational instability and that, in the specific case of liver cancer, this instability is mostly due to exposure to external mutagens.

Results

Human BSEP-HCCs do not accumulate mutations in cancer genes. To map the somatic mutations acquired in BSEP-HCC, we sequenced the whole exomes of six HCCs and corresponding background livers (Table 1 and Supplementary Data 1). After exome capture, Illumina sequencing and alignment of sequence reads (Supplementary Data 1), we identified a total of 44 single-nucleotide variants (SNVs) and 8 small insertions and deletions (indels) that were present in the six cancer exomes but not in the corresponding background livers (see Methods, and Table 1). The orthogonal validation of representative mutations assessed 93% specificity of our variant calling strategy (see Supplementary Methods). No mutation was shared between any two lesions, and each exome contained on average 0.05 somatic mutations per Mbp (Table 1). This mutation frequency was lower than the average mutation frequency of other human HCCs and of other

Table 1 | Somatic mutations and copy-number alterations in human BSEP-associated liver cancers.

Lesion ID	Gender	Age (years)	Tumour content (%)	Somatic SNVs	Non-silent SNVs	Somatic indels	Amplified genes	Deleted genes	LOH genes
175	M	1.6	90	7	3	0	10,688	1	4,964
7860	F	2.6	90	5	2	0	13,594	1,248	891
23836	M	1.3	90	5	1	0	12,450	19	2,847
HB4R	F	8.6	70	25	9	7	5,598	2,575	7,124
1790	M	11.7	60	1	0	0	8,601	0	158
2896	F	1.3	50	1	0	1	9,687	244	3,702
UKT	M	1.3	40	NA	NA	NA	3,801	258	0
Total	—	—	—	44	15	8	18,428	3,628	9,757

F, female; LOH, loss of heterozygosity; M, male; NA, not available; SNV, single-nucleotide variation.
Tumour content is expressed as percentage of tumour tissue in the inspected sample on histological analysis. Non-silent SNVs indicate mutations leading to protein modifications. None of the identified indels introduced a frameshift. Total refers to modifications found in at least one sample. The higher number of mutations found in sample HB4R compared with other samples probably reflects that this was a relapse and that the patient had been treated with chemotherapy.

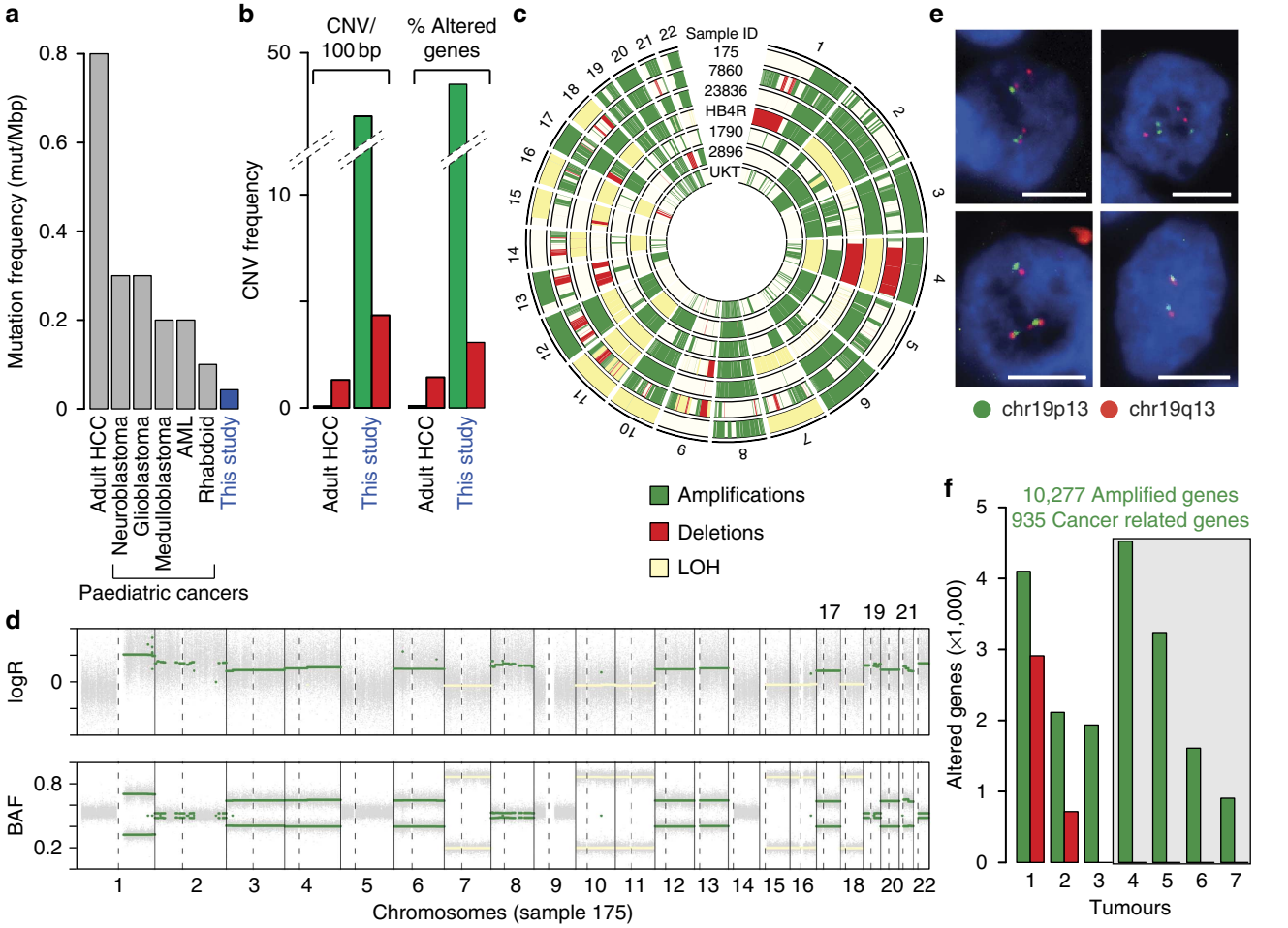


Figure 1 | Somatic genomic alterations occurring in BSEP-HCCs. (**a**) Average non-silent mutations frequency in the six BSEP-HCCs as compared with adult HCC and paediatric cancers[17]. (**b**) Average CNV frequency in the seven BSEP-HCCs as compared with adult HCC[3,6,7]. Frequency was measured as CNVs per 100 base pairs and as percentage of altered genes among total human genes. Detailed comparison with each adult HCC screening is reported as Supplementary Fig. 1. (**c**) Circos plot reporting amplifications, deletions and LOH in all chromosomes of the seven BSEP-HCCs. (**d**) Copy-number profile of sample 175, as detected from genome wide SNP array. The upper and lower panels show the log ratio of intensity and the B-allele frequency (BAF) of germline heterozygous SNPs, respectively. Amplified, deleted and LOH segments are shown in each chromosome. Dotted lines identify centromeres. (**e**) Amplification of chromosome 19 in patient 23836 as validated by FISH. Shown are four representative hepatocyte nuclei (blue, DAPI staining) from a paraffin-embedded tumour sample. Probes for chr19p13 (green) and for chr19q13 (orange) were used. Scale bars = 20 μm. (**f**) Amplified (green) and deleted (red) genes in the seven BSEP-HCCs, grouped according to the number of samples in which they were altered.

paediatric cancers[17] (Fig. 1a). To exclude the possibility that we did not detect mutations due to poor sensitivity, we measured the false negative rate of the variant calling method at different percentages of tumour content (see Supplementary Methods and Supplementary Table 1). We estimated 100% sensitivity in detecting somatic variants in lesions with >40% tumour content.

Of the 44 somatic SNVs, 15 led to non-silent modifications in 15 genes (Supplementary Data 2). Inspection of these genes pinpointed no strong candidates as drivers of BSEP-HCC. First, none of the mutated genes was a known driver gene in HCC or in any other human cancers[18,19]. Second, some of them were known passengers that recurrently mutate in several cancer types[20,21]. Finally, most of these genes are either poorly or not expressed in the human liver (Supplementary Data 2), thus indicating that they are probably not functional in this tissue.

Overall, our results showed low levels of mutational instability in BSEP-HCC, thus suggesting that the acquisition of mutational instability is not the driving force for the development of this type of liver cancer.

Massive gene amplification occurs in human BSEP-HCCs. We used genome-wide SNP arrays to investigate the occurrence of copy-number variations (CNVs) in seven BSEP-HCCs, including all tumours screened for point mutations and one additional lesion (Table 1). Compared with adult HCCs, BSEP-HCCs showed higher frequency of CNVs (Fig. 1b and Supplementary Fig. 1), with a total of 18,428, 3,628 and 9,757 genes that underwent amplification, deletion or loss of heterozygosity, respectively, in at least one of the seven samples (Table 1 and Supplementary Data 3). Low-level copy-number gains (on average four copies per aberrant region, Supplementary Fig. 2, Supplementary Data 3) were the most pervasive alterations, with around 38% of the total genome amplified in each sample (Supplementary Fig. 3A). We further analysed whether copy-number alterations were focal or arm-level events[22] and observed that deletions were mostly focal and sample specific, except for sample HB4R that had three arm-level deletions involving chromosomes 1p, 4q and 17p (Supplementary Fig. 3B). Copy-number gains were instead either arm-level (Supplementary Fig. 3B) or multiple focal events that led to the amplification of

entire chromosomes, most notably chromosomes 8, 19 and 20 in the majority of lesions (Fig. 1c,d and Supplementary Fig. 3A). To validate these results, we performed fluorescence *in situ* hybridization (FISH) with probes located on both arms of chromosome 19 and confirmed the amplification of this chromosome in sample 23836 (Fig. 1e). Multiple amplifications likely occurred via step-wise DNA rearrangements rather than through one-shot catastrophic events, because in most samples we did not detect any sign of chromothripsis. The only exception was sample UKT where several consecutive oscillations between two copy-number states in chromosomes 4 and 6 (Supplementary Fig. 4A,B, Supplementary Data 3) may suggest the occurrence of a one-shot catastrophic rearrangement event[23,24].

To find possible cancer drivers, we focused on genes that were recurrently altered in the majority of samples and in particular on 935 known cancer genes that were amplified in the genome of at least four of the seven lesions (Fig. 1f). Pathway enrichment analysis of these genes highlighted three top-scoring pathways, namely the MAPK, the ErbB and the PI3K/Akt pathways (corrected $P = 3 \times 10^{-06}$, 9×10^{-06} and 2×10^{-05}, respectively, hypergeometric test, Supplementary Data 4). These pathways form a complex and interconnected signalling network[25,26], and their activation is a known driver event in some types of liver cancer[27–29].

The results of the copy-number analysis showed that BSEP-HCCs are characterized by a pervasive occurrence of chromosomal rearrangements that lead to massive gene amplification. Recurrent events involve the amplifications of signalling genes, thus suggesting that the alteration of signalling pathways may be involved in the development and progression of this type of HCC.

***Mdr2*-KO HCCs resemble the genomic landscape of BSEP-HCCs**. To validate the potential contribution of the genomic modifications in BSEP-HCC, we profiled the cancer genomes of *Mdr2*-KO mice, which, similarly to human BSEP-deficient patients, develop HCC due to impairment of bile secretion[12,15]. We sequenced the exomes of nine HCCs extracted from the livers of seven *Mdr2*-KO mice using the kidney of one of them as a reference (Table 2). We applied the same procedures as those used with human samples for target enrichment, Illumina sequencing and variant calling (Supplementary Data 5) and identified a total of 118 somatic SNVs and no indels (Table 2). Also in this case, we confirmed >93% specificity by orthogonal validation of randomly selected variants (see Supplementary Methods). None of the 118 SNVs was shared between any two tumours, and 60 of them led to modifications in 60 proteins (Supplementary Data 6). As with BSEP-HCCs, no mutated gene was a known driver of HCC or of other cancers[18,19], and only a few of them were expressed in the liver (Supplementary Data 6). We further sequenced the coding exons of 866 mouse orthologues of human cancer genes (Supplementary Data 7) in four additional *Mdr2*-KO HCCs, using the normal liver as a reference. In this case, we increased the depth of sequencing coverage to further exclude that mutations might have been missed because of high intratumoral heterogeneity (Supplementary Data 5). Again we found no somatic mutations and no small indels in any cancer genes in any of the four samples (Table 2).

To assess whether the massive copy-number alteration observed in BSEP-HCC also occurred in mouse tumours, we developed a novel method to investigate CNVs directly from targeted re-sequencing data. Our procedure was based on the comparison of normalized gene coverage between tumour and reference that led to the identification of tumour-specific copy-number gains and losses (see Supplementary Methods). In the 13 mouse HCCs, we identified a total of 2,510 altered genes, almost all of which (2,507) were amplified (Table 2 and Supplementary Data 8). Validation of 10 randomly selected amplified genes using TaqMan copy-number assay estimated 70% sensitivity, 93% specificity and 84% accuracy of the method to call amplifications (Supplementary Table 2). We used genes on chromosome X to assess the performances of our method to detect deletions and estimated 91% sensitivity and 85% accuracy (see Supplementary Methods and Supplementary Table 3). We further sequenced the whole genomes of two late-stage *Mdr2*-KO HCCs from two different mice, using the corresponding kidneys as matched references. We again observed an overall higher occurrence of amplifications than deletions, with a total of 1,074 amplified and 117 deleted genes in the two genomes (Supplementary Data 8). Since one of the two tumours (ID: 60400/1) had been also used for exome sequencing, we further assessed the performance of CNV detection from targeted re-sequencing data. By far most of the amplified genes detected in the whole exome were also found in the whole genome (85%, Supplementary Data 8), thus confirming the reliability of the method.

Mdr2-KO tumours have already been shown to have high degrees of chromosomal instability. In agreement with our results, HCCs from *Mdr2*-KO mice that underwent partial

Table 2 | Somatic mutations and copy-number alterations in *Mdr2*-KO mice.

Lesion ID	Gender	Age (months)	Size (cm)	Tumour content	Sequenced regions	Somatic SNVs	Non-silent SNVs	Amplified genes	Deleted genes
51509/1	M	16	1.1	20%	Whole exome	8	5	59	0
60400/2	F	13	1.4	40%		8	3	113	0
218/1	M	15	1	50%		5	2	298	1
52686/1	F	15	0.7	50%		8	4	15*	2
58853/3	M	15	1.7	60%		20	8	631	0
60400/1	F	13	0.9	60%		9	3	455	0
58163/3	M	15	3	70%		17	6	333	1
58163/4	M	15	3	70%		39	27	625	0
215/1	M	14	1.8	80%		4	2	562	0
54913/10	F	10	0.1	NA	866 Cancer genes	0	0	49	0
54913/8	F	10	0.5	NA		0	0	10	0
55481/10	F	10	0.3	NA		0	0	17	0
55484/4	F	10	3	30%		0	0	41	0
Total	—	—	—	—	—	118	60	2,507	4

F, female; M, male; NA, not available; SNV, single-nucleotide variation.
Nodules with insufficient amount of tissue for histologic inspection where defined as NA.
*TaqMan copy-number assay assessed a high number of false negatives for this sample, thus suggesting an overall underestimation of gene amplifications.

hepatectomy show pervasive amplifications, no detectable deletions and recurrent amplifications in chromosomes 5, 8 and 18 (ref. 30). Moreover, genes in the 20 Mbps around the centromere of chromosome 8 are upregulated in *Mdr2*-KO HCCs[16]. Interestingly, genes in this region were among the most recurrently amplified in our samples (see below and Supplementary Data 8).

The results of the genomic profiling of *Mdr2*-KO HCCs showed that, similarly to human BSEP-HCCs, these tumours are not prone to accumulate somatic point mutations and small indels. Instead, they show pervasive chromosomal instability with overwhelming prevalence of gene amplification.

Somatic CNVs accumulate during *Mdr2*-KO HCC progression. We noticed that human BSEP-HCCs with high tumour content (90%) accumulated more amplifications than those with lower tumour content (Table 1). Interestingly, a positive correlation between tumour size and fold change of the amplified regions was already reported in HCCs from *Mdr2*-KO mice that underwent partial hepatectomy[30]. We therefore checked whether a similar signal was detectable in *Mdr2*-KO HCCs of our cohort. We indeed confirmed a positive correlation between the number of amplified genes and the HCC content (Pearson's correlation coefficient = 0.78, $N = 10$, Fig. 2a). Moreover, bigger lesions showed significantly more amplified genes than smaller lesions ($P = 0.03$, Wilcoxon test, $N = 10$, Fig. 2b). These results suggested that amplifications tend to occur in larger and more advanced lesions. Furthermore, amplifications preferentially accumulated near the centromeres (Fig. 2c) of mouse chromosomes, which are known hotspots for mitotic recombination[31].

In the two *Mdr2*-KO HCC genomes, we identified inverted translocations involving chromosomes 8 and 14 in one tumour (ID: 218/3, Fig. 2d) and chromosomes 8 and 19 in the other (ID: 60400/1, Fig. 2e). Interestingly, mouse chromosome 8 is the ortholog of human chromosomes 8 and 19, which are the most recurrently amplified chromosomes in human BSEP-HCCs (Supplementary Fig. 3A). Through the analysis of discordantly aligned read pairs, we were able to map the two breakpoints at base pair resolution and both rearrangements were confirmed with PCR amplification and Sanger sequencing (Fig. 2d,e). In one of the two tumours (ID: 60400/1), we counted 10 consecutive disomic and trisomic copy-number states in the region of chromosome 19 involved in the inverted translocation (Fig. 2e and Supplementary Fig. 3C). Therefore, again similarly to the human samples, we found possible indications that one-off catastrophic events could be responsible for the acquisition of at least some of the genomic rearrangements in this liver cancer type.

In summary, the CNV analysis of *Mdr2*-KO HCCs showed that they undergo frequent genomic rearrangements that lead to gene amplifications. Moreover, the observation that CNVs tend to accumulate in large lesions with high HCC content suggests that their putative driver role is exerted in tumour progression rather than in tumour initiation.

JNK is deregulated in *Mdr2*-KO HCCs. Given the similarities in genomic landscapes of acquired alterations between BSEP-HCCs and *Mdr2*-KO HCCs, we performed pathway enrichment analysis on 27 genes that were recurrently amplified in both human and mouse cancers (Fig. 3a, Supplementary Data 8 and Supplementary Table 4). Again we found that the MAPK signalling cascade was among the top scoring pathways (corrected $P = 2 \times 10^{-02}$, hypergeometric test, see Supplementary Methods). In particular, *Map2k7*, encoding the mitogen-activated protein-kinase kinase, was amplified in >70% human and mouse HCCs (5 out of 7 and 10 out of 14, respectively). To better quantify the frequency of *Map2k7* amplification, we screened 35 additional tumours from 16 distinct mice using TaqMan copy-number assay. *Map2k7* was amplified in 14 of the 49 *Mdr2*-KO nodules that were analysed overall (29%, Supplementary Data 9), which was a significantly higher proportion than expected by chance ($P = 6 \times 10^{-04}$, binomial test). When only nodules with high HCC content were considered, the enrichment became even stronger, with 58% of nodules with ≥40% HCC showing *Map2k7* amplification ($P = 9 \times 10^{-05}$, binomial test). This result further supports the general observation that gene amplification tends to occur preferentially in lesions with high tumour content (Fig. 2a). To investigate whether the additional copies of *Map2k7* directly impinge on gene expression, we measured the mRNA levels in tumours where the gene was amplified as compared with tumours where it was not amplified, with *Mdr2*-KO inflamed livers and with age-matched healthy livers from *Mdr2*-wild type mice. Tumours with *Map2k7* amplifications showed significant *Map2k7* overexpression compared with all other groups (Fig. 3b), thus indicating that gene amplification led to increased expression. Map2k7 specifically regulates the c-Jun NH(2)-terminal kinases (JNKs)[32–34], which are mainly activated mainly by pro-inflammatory cytokines and environmental stress[35]. Deregulation of JNKs has been already reported to play cell- and stage-dependent roles during HCC development[36–40]. Interestingly, upstream and downstream direct JNK interactors, as well as JNKs themselves, were altered in several human and mouse samples (Fig. 3c).

These data showed that gene amplifications occurring in *Mdr2*-KO tumours preferentially hit signalling genes, most notably JNK direct interactors or upstream activators, which may have a driver role in triggering liver tumour progression.

JNK inhibition arrests carcinoma progression in *Mdr2*-KO mice. We set out to investigate whether JNK inhibition might interfere with liver cancer progression by treating *Mdr2*-KO mice with SP600125, a synthetic polyaromatic chemical that directly inhibits the JNK kinases[41–46]. We randomized 23 *Mdr2*-KO mice to receive either SP600125 or vehicle only (12 and 11 mice, respectively, Supplementary Data 10). After 3 weeks of treatment, mice were killed and the tumours from the two cohorts were compared in terms of *Map2k7* amplification, nodule number, size, histology and tumour content. We found a significantly lower proportion of lesions from treated mice (5 out of 36, 14%) with *Map2k7* amplification when compared with lesions from the untreated cohort (13 out of 35, 37%, $P = 0.03$, Fisher's exact test, Supplementary Data 10). Thus, tumours with *Map2k7* amplification were more sensitive to JNK inhibition than those without *Map2k7* amplification, which explains their relative depletion after treatment. Indirectly, this result also indicated that the effects of SP600125 were mainly caused by its on-target activity on JNK. Moreover, although treated and untreated animals showed a comparable number of tumours per mouse, no mouse treated with the JNK inhibitor had nodules bigger than 20 mm, which instead represented ~20% of all lesions in the untreated group (Fig. 4a,b Supplementary Data 10). In a reciprocal manner, the proportion of nodules with diameters between 10 and 20 mm was significantly higher in treated mice than in the untreated group (Fig. 4a). Finally, nodules bigger than 10 mm showed significantly higher proportion of adenoma and lower proportion of adenocarcinoma in treated than in untreated mice (Fig. 4c,d). No difference was detectable in the histological composition of small lesions (diameter <10 mm). Similarly, treated mice showed an overall significant depletion in HCC, while purely adenomatous nodules were over-represented (Fig. 4e). Altogether, these data suggested that the drug blocks

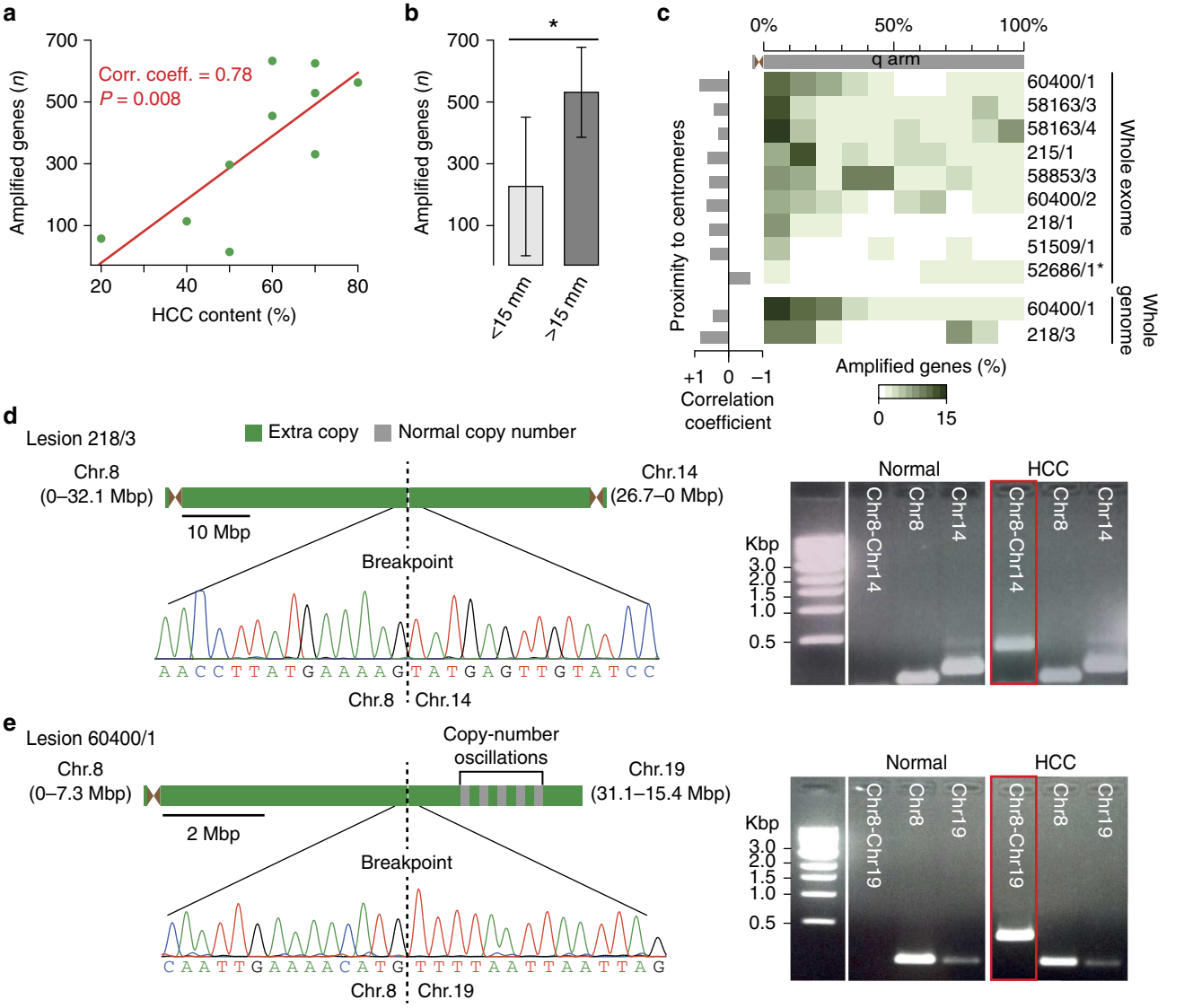


Figure 2 | Copy-number variations and complex structural rearrangements in *Mdr2*-KO HCCs. (**a**) Correlation between the number of amplified genes and HCC content in mouse tumours that underwent whole-exome and whole-genome sequencing. Correlation coefficients were measured using the Pearson correlation testing. (**b**) Comparison between number of altered genes in small (<15mm) and big (>15mm) lesions. *$=P<0.05$, Wilcoxon test, $N=10$. Minimum and maximum number of altered genes in the two groups are shown. (**c**) Cumulative fractions of gene amplifications (amplified genes/total genes) in regions representing 10% of the q arm-length in all mouse chromosomes. In case of exome re-sequencing, the chromosome length was calculated as the region from the first to the last targeted base in the SureSelect XT Mouse All Exon kit (Agilent). Pearson correlation coefficients were calculated between the fraction of amplified genes in each region and the proximity to the centromere of each chromosome. *52686/1 was the only tumour with a negative correlation, likely due to overall CNV underestimation in this sample (see also Table 2). (**d**) Inverted translocations between chromosomes 8 and 14 of lesion 218/3 and (**e**) chromosomes 8 and 19 of lesion 60400/1. Through the analysis of discordant sequencing read pairs, breakpoints of both rearrangements were mapped at base pair resolutions and confirmed by PCR amplification and Sanger sequencing of breakpoint regions. In sample 60400/1, 10 consecutive copy-number oscillations at chromosome 19 were detected in proximity of the rearrangement with chromosome 8. The genomic coordinates of amplified regions in each chromosome are shown in parentheses.

tumour progression towards bigger lesions with higher HCC content, thus supporting the role of JNK deregulation in progression more than in initiation of *Mdr2*-KO tumours. They were also consistent with the results of the CNV analysis, and specifically with the tendency of gene amplification, and of *Map2k7* amplification in particular, to occur in large lesions with high HCC content.

Discussion

In this study, we profiled the genomes of aetiologically related mouse and human HCCs to identify the genomic changes acquired in response to chronic exposure to non-neutralized bile acids and in the absence of exogenous direct (viruses) or indirect (alcohol) mutagens. Despite the small sample size that we analysed, our screenings showed a consistent genomic signature within and between species. In both human and mouse, cancer genomes accumulated massive copy-number gain in contrast to very few somatic SNVs or small indels. Such a genomic signature is remarkably different from that of the other HCCs previously sequenced, which acquire mutational instability and tend to accumulate gene deletions rather than amplifications[3]. These findings confirm the genetic heterogeneity of liver cancers caused by different aetiological agents and, at the same time, the

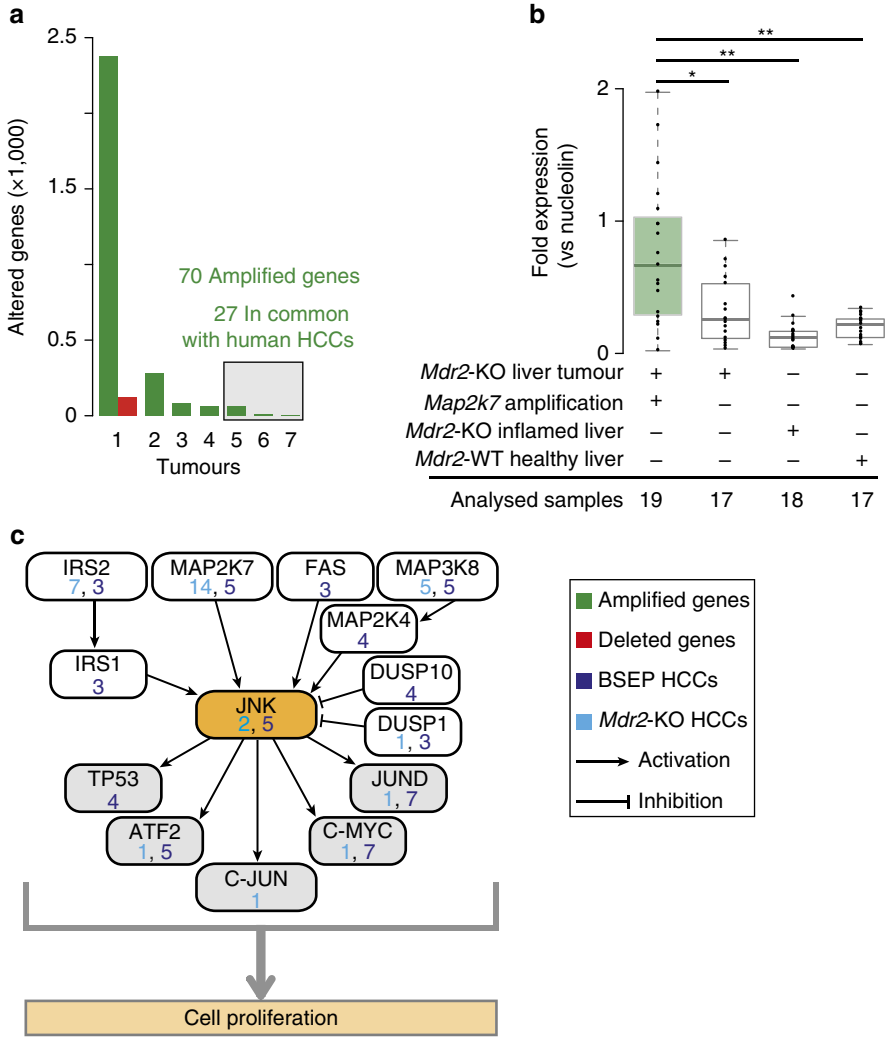


Figure 3 | Recurrent gene amplifications of JNK interactors in BSEP- and *Mdr2*-KO HCCs. (**a**) Amplified (green) and deleted (red) genes in all sequenced *Mdr2*-KO HCCs, grouped according to the number of samples in which they were altered. No gene was modified in more than seven tumours. (**b**) *Map2k7* expression measured by qPCR in *Mdr2*-KO nodules where the gene was amplified, in nodules with no amplification, in *Mdr2*-KO inflamed livers and in age-matched *Mdr2*-WT livers. Dots represent the different samples in each group. * = $P<0.05$, ** = $P<0.005$, Wilcoxon test, $N=71$. Maximum, minimum and median are shown for each distribution. (**c**) JNK interactors that are altered in *Mdr2*-KO and BSEP-HCCs. Number of lesions with the amplified gene are reported for mouse (light blue) and human (deep blue).

remarkable analogy among human and mouse tumours with similar aetiopathogenesis.

The detailed analysis of *Mdr2*-KO mouse cancer genomes revealed that copy-number gains tend to cluster in genomic regions with high mitotic recombination rates, such as centromeres and telomeres (Fig. 2c). These regions are fragile sites prone to rearrangements[47] and are associated with loss of heterozygosity[48] and replication stress[49–51]. Chronic inflammation induces the production of reactive oxygen species that may lead to oxidative DNA damage, thus possibly explaining why genes associated with chromosomal instability are upregulated in the liver of *Mdr2*-KO mice[30,52]. It is therefore tempting to speculate that recombination hotspots are directly involved in cancer genomic instability in this tumour type, in the absence of external causes of DNA damage. This is also compatible with the role of inflammation in inducing a hypoxic microenvironment that favours chromosomal instability[53–55].

Frequent amplifications of JNK activators, most notably the recurrent amplification of *Map2k7*, were found in the majority of mouse and human cancers, suggesting that these modifications may be the driver events in this liver cancer type. Since amplifications of JNK activators preferentially occur at late stages of HCC development, the deregulation of this pathway is likely to favour cancer progression rather than disease initiation. JNK is involved in several physiological and pathological processes including cell proliferation, differentiation, apoptosis and tumorigenesis[35]. Its activation has already been implicated in the development of liver cancer. For example, neither *c-Jun* nor *JNK1*-deficient mice develop HCC after exposure to mutagens[37,39,40,56]. In addition, liver-specific deletion of *Mapk14*, a negative regulator of *Map2k7*, leads to JNK hyperactivation and HCC development in the mouse[38]. Interestingly, JNK deficiency reduces the onset of inflammation and tumorigenesis when occurring in both hepatocytes and non-parenchymal cells, but it is linked only to increased tumour size when limited to hepatocytes[57]. This hints at an oncogenic role of JNK in non-parenchymal cells where it likely promotes an inflammatory environment that favours transformation and/or tumour progression. Our data support such a scenario, indicating that JNK amplification leads to its deregulation and favours tumour progression in BSEP- and *Mdr2*-KO HCCs. In contrast, the JNK pathway is not recurrently amplified in virus-,

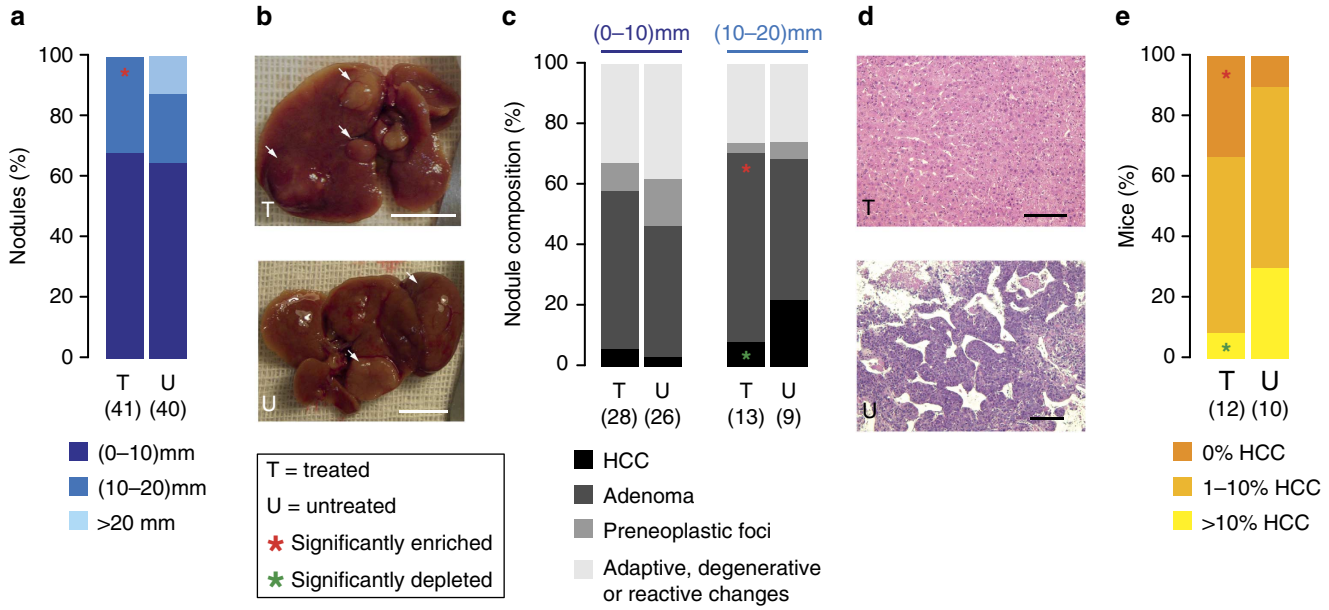


Figure 4 | Effect of JNK inhibition on *Mdr2*-KO HCC progression. (**a**) Size differences in nodules from treated and untreated *Mdr2*-KO mouse groups. Nodules from treated mice were significantly enriched in 10–20 mm lesions, but had no lesions >20 mm. The latter represented ~20% of nodules in untreated mice. (**b**) Representative images of livers from a treated and an untreated mouse. Arrows indicate the nodules. Scale bar = 1cm (**c**) Histological composition of nodules from treated and untreated mice. Nodules in the two cohorts of mice were divided into two groups by size (<10 mm and >10 mm). (**d**) Representative photomicrographs histologic sections of HCC and adenoma from treated and untreated livers, respectively (hematoxylin/eosin). Scale bar = 150 µm. (**e**) Cumulative tumour content in treated and untreated mice. Tumour content was measured as a percentage of HCC in each nodule. Nodules with HCC fraction >10%, <10% and with no HCC were compared between treated and untreated mice. In all analyses, differences were assessed using Fisher's exact test. The number of nodules or mice in the two groups are reported in parentheses.

alcohol- and other risk factor-associated HCCs[3–8]. This further highlights that different disease aetiologies have distinct impacts on tumour genomics, which in turn may lead to different molecular mechanisms of tumour initiation and progression. Since we observed that pharmacological inhibition of JNK impairs the adenoma-to-carcinoma progression *in vivo*, JNK inhibition may be useful to block HCC onset in BSEP deficiency patients waiting for liver transplantation. In this regard, it is important to notice that the drug used to block JNK (SP600125) has been reported to exert secondary effects on targets other than JNKs[58,59]. However, the fact that tumours with *Map2k7* amplification were significantly depleted on treatment suggests that the effects of SP600125 were mainly accounted for by the ability to inhibit JNK.

In conclusion, this study demonstrates that intrahepatic cholestasis leading to hepatocyte exposure to bile acids and chronic inflammation generates a unique and distinctive signature of genomic changes that can be clearly distinguished from those caused by viruses and other external factors. It will be interesting to determine whether and to what extent similar genomic changes represent a general feature of tumours arising from other chronically inflamed tissues.

Methods

Human sample description and DNA extraction. Samples used in the study were obtained from frozen or formalin-fixed paraffin-embedded (FFPE) material from seven children diagnosed with BSEP-HCC, with parental written consent (Supplementary Data 1). The protocol for use of human tissues was approved by the review board of the corresponding hospital of provenience (French Institute of Medical Research and Health IRB Number 11-047; UK Integrated Research Application System ID: 103273, REC reference:12/WA/0282; Italian Ministry of Health, statement 61, 19/12/1986 N.900.2/ Ag 464/260; Ethical Review Board of the University Hospital Tübingen, Ref.no, 27/2008B01). All specimens were obtained at native-liver hepatectomy during transplantation. The background liver in all patients exhibited parenchymal rather than portal-tract cholestasis, with BSEP expression detectable in none. Some patients had frank cirrhosis, others only fibrosis, which varied in degree from patient to patient (Supplementary Data 1). Samples 7860 175, 1790, 2896 and UKT came from single-unencapsulated masses, while sample 23836 derived from one of several HCC within a single liver. Sample HB4R was a relapse that developed within allograft liver 6 years after transplantation. The patient was treated with chemotherapy before relapse and surgical resection. In all samples, non-neoplastic liver tissues from the same patients were used as matching background references.

Genomic DNA was extracted from each tumour from matched background liver tissue using the DNeasy Blood and Tissue Kit (Qiagen) for frozen samples and with the AllPrep DNA/RNA FFPE Mini Kit (Qiagen) for FFPE blocks.

Exome sequencing, variant calling and mutation validation. Target capture was done on six human tumour and reference samples (Supplementary Data 1) using the SureSelect XT Human All Exon V4 kit (Agilent) targeting 20,965 human genes, following the manufacturer's protocol with minor modifications. Sample UKT was excluded from whole-exome sequencing because of the low tumour content (see Supplementary Methods). Around 3 µg of genomic DNA was sheared using an Adaptive Focused Acoustics technology (Covaris). After library preparation with an Illumina DNA Sample Prep Kit, 200 bp fragments were selected using the Agencourt AMPure PCR Purification system (Beckman Coulter). Fragments were further amplified with 5 to 7 cycles of PCR and 500 ng was hybridized with the bait library. DNA capture was followed by paired-read cluster generation on the Cluster Station (Illumina). Libraries were sequenced using one-half lane of Illumina HiSeq2000 per sample, with 76 bp or 101 bp paired-end protocol, except for the tumoral sample of patient 7860, where one entire lane was used due to high levels of DNA degradation (Supplementary Data 1).

Paired-end sequencing reads from each tumour and reference were mapped to the human genome (GRCh37/hg19) using Novoalign (http://novocraft.com). At most three mismatches per read were allowed and duplicated reads were removed using rmdup of SAMtools[60]. All reads uniquely mapping within 75–100 bp of the targeted regions were considered on target and retained for further analysis (Supplementary Data 1). SNVs and indels were identified using SAMtools[60] and VarScan 2 (ref. 61) and retained if covered by at least 10 reads and with frequency ≥20%. Somatic mutations and indels were identified as tumour-specific mutations with coverage ≥5x, frequency <10% in the reference, and not present in dbSNP build 137 (MAF >1%). All 44 SNVs and 8 indels were retained after manual inspection, and 14 non-silent SNVs underwent orthogonal validation (see Supplementary Methods).

Functional annotation of mutated genes. The list of genes affected by mutations was intersected with the genes known to be recurrently mutated in HCC (*TP53*,

CTNNB1, *ARID1A*, *ARID2*, *AXIN1*, *RPS6KA3*, *VCAM1*, *CDK14*, *TERT*, *MLL4*, *CCNE1*)[18] and with the list of 537 genes known to have a causative role in human cancer[19].

Expression levels of mutated genes in the normal liver were inferred from publicly available data[62,63]. Starting from the raw CEL files of the two experiments, data were normalized and analysed using the MAS5 algorithm. The expression level for each gene in the liver was calculated as the mean value among all gene probes with detection P-value <0.05. If all probes of a gene had P-value >0.05, the gene was considered not expressed. The normalized expression level was then measured as the gene expression level over the median expression of all genes in the liver. Genes with expression higher than the median were considered to be high expressed, while all genes with expression lower than the median were defined as low expressed.

Recurrently mutated passenger genes were retrieved from literature (refs 20,21 and http://bio.ieo.eu/ncg/).

SNP array and copy-number calling in the human samples. Genomic DNA extracted from FFPE samples and from frozen samples was processed according to the Infinium HD assay ultra manual. DNA from FFPE samples was restored before SNP array processing according to the Infinium HD FFPE restore protocol. All seven human tumours and matched background livers were assayed using Illumina HumanOmniExpress-12 v1.0, and image data were scanned using a BeadArray reader. Intensity and genotype data were extracted for CNV analysis after normalizing raw fluorescence signals using Illumina Genome Studio v2011.1. CNV analysis was performed using ASCAT (version 2.1), which takes into consideration aneuploidy and non-aberrant cell admixture present in each tumour sample[64] (see Supplementary Methods).

For six tumours with whole-exome sequencing data, frequency distributions of the germline heterozygous SNPs were integrated with the SNP array results to identify high confidence CNV regions (see Supplementary Methods). To identify altered genes, the genomic coordinates of the aberrant regions in each sample were intersected with those of 20,965 human genes of the SureSelect XT Human All Exon V4 kit (Agilent). A gene was considered as modified if $\geq$80% of its length was contained in an aberrant region.

Fluorescence *in situ* hybridization. Validation of amplification of chromosome 19 in sample 23836 was performed by two-colour fluorescence *in situ* hybridization (FISH) using a Vysis LSI 19q13 SpectrumOrange/19p13 SpectrumGreen probe (Abbott), according to the manufacturer's instructions. Two-micrometre FFPE slides from the tumour and background liver of patient 23286 were deparaffinized in xylene, washed in 100% ethanol, incubated in 1 × SSC (0.3 M sodium chloride, 0.03 M sodium citrate) pH 6.0 at 80 °C for 20 min for demasking and digested with pepsin (0.5 mg ml^{-1} in 0.2 N HCl, pH 1.0; Protease and Protease Buffer II, Abbott) for 17 min at 37 °C. Samples were then washed in 2 × SSC, dehydrated in 70, 95 and 100% ethanol and air dried. Ten microlitres of probe was directly applied onto each slide and topped with a coverglass that was then sealed with rubber cement. Slides were placed in a HYBrite (Abbott), and the probe was left to denature 1 min at 85 °C, followed by an overnight hybridization at 37°C. Coverglasses were then removed and slides were washed twice in 2 × SSC with 0.1% NP-40 at RT, once in 0.4 × SSC with 0.3% NP-40 at 73 °C and once again in 2 × SSC with 0.1% NP-40 at RT. After counterstaining with DAPI (Sigma), FISH signals were scored with an Olympus *BX61* upright microscope, using a ×100 objective.

Gene enrichment and pathway analysis. Pathway enrichment was done with ConsensusPathDB[65]. A total of 935 human cancer genes amplified in at least four BSEP-HCCs, and 27 genes that were amplified in the majority of human and mouse HCCs, were compared with the pathway-base gene set composed of 10,529 pathway-associated genes. P-values were calculated with the hypergeometric test based on the number of pathway components present in both the amplified cancer gene set and the pathway-base gene set. The resulting P-values were then corrected for multiple testing using false discovery rate.

Mouse description and DNA extraction. Experiments involving mice have been done in accordance with the Italian Laws (D.L.vo 116/92) and mice have been housed according to the guidelines of the European Commission Recommendation 2007/526/EC—June 18, 2007. The project has been notified to the Italian Ministry of Health (project n. 106/11). Founders of the FVB.129P2-Abcb4^{tm1Bor}/J (*Mdr2*-KO, stock number: 002539) and FVB/NJ (*Mdr2* wild type, stock number:001800) mice were purchased from The Jackson Laboratory. Colonies of both strains were maintained under specific pathogen-free conditions. Adenoma and HCC nodules from *Mdr2*-KO mice were snap frozen for DNA/RNA extraction or fixed in formalin for histological analysis. Initial pathological stage (inflammation) DNA/RNA extraction was carried out on purified populations of hepatocytes obtained via collagenase liver perfusion, using a two-step protocol[66]. The normal livers or kidneys were collected and frozen to be used as reference. Frozen tissue samples were homogenized with a GentleMACS Dissociator (Miltenyi Biotec) before column extraction. Genomic DNA was extracted using the DNeasy Blood and Tissue Kit (Qiagen) according to the manufacturer's protocol for all samples and matching wild-type tissue.

Mouse histology. All analysed samples were inspected by a mouse pathologist. Tumour growth in *Mdr2*-KO livers is multicentric; grossly detectable masses are often very heterogeneous resulting from the collision of multiple contiguous hepatocellular proliferations with different histologic features and grades. In addition, there is a tendency for hepatocellular carcinoma to develop within an adenoma (as foci of tumour progression). In this context, local invasion (one of the most reliable indicators of tumour malignancy) is difficult to assess and carcinoma diagnosis relies entirely on the recognition of clear features of architectural or cytologic atypia. Given these peculiar characteristics of tumour growth in *Mdr2*-KO mice, the histological composition of grossly detectable hepatic nodules was semiquantitatively determined based on reported classification criteria[67].

Selection of the 866 mouse orthologues of human cancer genes. A collection of 2,061 human cancer genes was selected from the Cancer Gene Census[19], COSMIC (http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/) and high-throughput cancer mutational screenings (http://bio.ieo.eu/ncg/). The 1,753 mouse orthologues of these genes were identified using eggNOG[68] and MGI (http://www.informatics.jax.org), and only the 866 with RefSeq entries were selected (Supplementary Data 7). The SureSelect Custom kit (Agilent) was designed to capture 15,067 exons of the 866 genes, for a total of 2.7 Mbp of DNA. Exons shorter than 60 bp (except for those with mutations in COSMIC), sequence repeats, segmental duplications, PAR regions and gaps were excluded. Finally, only regions with GC content ranging from 30 to 65% were selected to optimize the capture efficiency.

Whole-exome and whole-genome sequencing of mouse samples. The SureSelect custom kit was used for the 866 selected genes and the SureSelect XT Mouse All Exon kit (Agilent) was used to target 21,543 mouse genes following the manufacturer's protocol with slight modifications. In brief, around 3 µg of genomic DNA was sheared using an ultrasonic disruptor (Bioruptor, Diagenode) or using an Adaptive Focused Acoustics technology (Covaris). After library preparation with the Illumina DNA Sample Prep Kit, 200–250 bp fragments were selected and purified by gel extraction, or using the minelute PCR purification kit (Qiagen), or using the Agencourt AMPure PCR Purification system (Beckman Coulter). Fragments were further amplified with 10 cycles of PCR and 500 ng was hybridized with each bait library. DNA capture was followed by single- or paired-read cluster generation on the Cluster Station (Illumina). The libraries obtained for the 866 genes were sequenced on the Genome Analyzer IIx with the 76 single-end protocol, using one lane for each tumour sample or matching normal sample. The libraries obtained for the whole exomes were sequenced using one-half lane of Illumina HiSeq2000 per sample, with the 101 bp paired-end protocol (Supplementary Data 5). For whole-genome sequencing, around 1 µg of mouse genomic DNA was sheared in 400–500 bp fragments using an ultrasonic disruptor (Bioruptor, Diagenode). Libraries were prepared with Illumina Paired-End DNA Sample Prep Kit. The libraries obtained were sequenced using one lane of Illumina HiSeq2000 per sample, with the 101 bp paired-end protocol (Supplementary Data 5).

CNVs and structural rearrangements in mouse samples. To detect CNVs on targeted re-sequencing data, we developed an in-house pipeline based on the difference in sequencing coverage between tumours and normal counterparts (see Supplementary Methods). Copy-number analysis on the whole-genome sequencing data was performed using CNVnator v. 0.2.5 (ref. 69) (see Supplementary Methods).

Structural rearrangements were inferred using PEMer[70] with slight modifications to adapt the method to Illumina sequencing. Paired-end insert size distribution was calculated to determine the expected insert size range. All mapped paired-end pairs that displayed either an insert size greater than expected or an unexpected orientation were selected. Of these, only discordant read pairs overlapping with or next to mapped CNV regions in the tumour were selected for manual inspection. Identified rearrangements between chromosomes 8 and 14 (ID: 218/3) and 8 and 19 (ID: 60400/1) were confirmed with PCR amplifications and Sanger sequencing.

Expression quantitation of *Map2k7*. Total RNA for qRT-PCR experiments was extracted from *Mdr2*-KO tumours, *Mdr2*-KO inflamed livers and age-matched *Mdr2*-WT healthy livers in Trizol (Invitrogen) using the RNeasy Mini Kit (Qiagen) according to the manufacturer's instructions. Total RNA (0.5 µg) was used for cDNA synthesis (using ImProm-II Reverse Transcriptase, Promega). Quantification was performed on Nanodrop, and quality was assessed on Bioanalyzer (Agilent). Expression analysis was carried out by qPCR on 1 µl of cDNA reverse-transcribed from 0.5 µg of total RNA. qPCR (SYBR Green, Applied Biosystems) analysis was performed on an Applied Biosystems 7500 Real-time PCR system. Values were normalized for nucleolin content.

Treatment with SP600125 JNK inhibitor. Twenty-three *Mdr2*-KO mice were randomly divided into two groups at the age of 13 to 14 months, when nodules are already formed[15]. One group of 12 mice was treated with an SP600125 (anthra[1,9-cd]pyrazol-6(2H)-one) (Calbiochem), and the other group of 11 mice with vehicle

Supplementary Data 10). Vehicle for SP600125, diluted in DMSO, was 40% polyethylene glycol (PEG, Sigma) in PBS. Treatments (60 mg per dose) were administered intraperitoneally three times a week for a total of 3 weeks. Mice were killed 1 week after the end of the treatment, and all grossly detectable nodules were counted, measured with a caliper and collected for DNA extraction and histological analysis.

References

1. Block, T. M., Mehta, A. S., Fimmel, C. J. & Jordan, R. Molecular viral oncology of hepatocellular carcinoma. *Oncogene* **22,** 5093–5107 (2003).
2. El-Serag, H. B. & Rudolph, K. L. Hepatocellular carcinoma: epidemiology and molecular carcinogenesis. *Gastroenterology* **132,** 2557–2576 (2007).
3. Guichard, C. *et al.* Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. *Nat. Genet.* **44,** 694–698 (2012).
4. Huang, J. *et al.* Exome sequencing of hepatitis B virus-associated hepatocellular carcinoma. *Nat. Genet.* **44,** 1117–1121 (2012).
5. Li, M. *et al.* Inactivating mutations of the chromatin remodeling gene ARID2 in hepatocellular carcinoma. *Nat. Genet.* **43,** 828–829 (2011).
6. Fujimoto, A. *et al.* Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. *Nat. Genet.* **44,** 760–764 (2012).
7. Totoki, Y. *et al.* High-resolution characterization of a hepatocellular carcinoma genome. *Nat. Genet.* **43,** 464–469 (2011).
8. Jiang, Z. *et al.* The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients. *Genome Res.* **22,** 593–601 (2012).
9. Brechot, C., Pourcel, C., Louise, A., Rain, B. & Tiollais, P. Presence of integrated hepatitis B virus DNA sequences in cellular DNA of human hepatocellular carcinoma. *Nature* **286,** 533–535 (1980).
10. Jacquemin, E. Progressive familial intrahepatic cholestasis. *Clin. Res. Hepatol. Gastroenterol.* **36,** S26–S35 (2012).
11. Knisely, A. S. *et al.* Hepatocellular carcinoma in ten children under five years of age with bile salt export pump deficiency. *Hepatology* **44,** 478–486 (2006).
12. Smit, J. J. *et al.* Homozygous disruption of the murine mdr2 P-glycoprotein gene leads to a complete absence of phospholipid from bile and to liver disease. *Cell* **75,** 451–462 (1993).
13. Fickert, P. *et al.* Regurgitation of bile acids from leaky bile ducts causes sclerosing cholangitis in Mdr2 (Abcb4) knockout mice. *Gastroenterology* **127,** 261–274 (2004).
14. Mauad, T. H. *et al.* Mice with homozygous disruption of the mdr2 P-glycoprotein gene. A novel animal model for studies of nonsuppurative inflammatory cholangitis and hepatocarcinogenesis. *Am. J. Pathol.* **145,** 1237–1245 (1994).
15. Pikarsky, E. *et al.* NF-kappaB functions as a tumour promoter in inflammation-associated cancer. *Nature* **431,** 461–466 (2004).
16. Katzenellenbogen, M. *et al.* Molecular mechanisms of liver carcinogenesis in the mdr2-knockout mice. *Mol. Cancer Res.* **5,** 1159–1170 (2007).
17. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339,** 1546–1558 (2013).
18. Zhang, Z. Genomic. landscape of liver cancer. *Nat. Genet.* **44,** 1075–1077 (2012).
19. Futreal, P. A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4,** 177–183 (2004).
20. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2013).
21. An, O. *et al.* NCG 4.0: the network of cancer genes in the era of massive mutational screenings of cancer genomes. *Database* **2014** (2014).
22. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463,** 899–905 (2010).
23. Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144,** 27–40 (2011).
24. Korbel, J. O. & Campbell, P. J. Criteria for inference of chromothripsis in cancer genomes. *Cell* **152,** 1226–1236 (2013).
25. Manning, B. D. & Cantley, L. C. AKT/PKB signalling: navigating downstream. *Cell* **129,** 1261–1274 (2007).
26. Raman, M., Chen, W. & Cobb, M. H. Differential regulation and properties of MAPKs. *Oncogene* **26,** 3100–3112 (2007).
27. Calvisi, D. F. *et al.* Ubiquitous activation of Ras and Jak/Stat pathways in human HCC. *Gastroenterology* **130,** 1117–1128 (2006).
28. Yea, S. *et al.* Ras promotes growth by alternative splicing-mediated inactivation of the KLF6 tumor suppressor in hepatocellular carcinoma. *Gastroenterology* **134,** 1521–1531 (2008).
29. Calvisi, D. F. *et al.* Increased lipogenesis, induced by AKT-mTORC1-RPS6 signaling, promotes development of human hepatocellular carcinoma. *Gastroenterology* **140,** 1071–1083 (2011).
30. Barash, H. *et al.* Accelerated carcinogenesis following liver regeneration is associated with chronic inflammation-induced double-strand DNA breaks. *Proc. Natl Acad. Sci. USA* **107,** 2207–2212 (2010).
31. Jaco, I., Canela, A., Vera, E. & Blasco, M. A. Centromere mitotic recombination in mammalian cells. *J. Cell Biol.* **181,** 885–892 (2008).
32. Davis, R. J. Signal transduction by the JNK group of MAP kinases. *Cell* **103,** 239–252 (2000).
33. Tournier, C. *et al.* MKK7 is an essential component of the JNK signal transduction pathway activated by proinflammatory cytokines. *Genes Dev.* **15,** 1419–1426 (2001).
34. Wada, T. *et al.* MKK7 couples stress signalling to G2/M cell-cycle progression and cellular senescence. *Nat. Cell Biol.* **6,** 215–226 (2004).
35. Weston, C. R. & Davis, R. J. The JNK signal transduction pathway. *Curr.Opin. Cell Biol.* **19,** 142–149 (2007).
36. Nikolaou, K. *et al.* Inactivation of the deubiquitinase CYLD in hepatocytes causes apoptosis, inflammation, fibrosis, and cancer. *Cancer Cell* **21,** 738–750 (2012).
37. Sakurai, T., Maeda, S., Chang, L. & Karin, M. Loss of hepatic NF-kappa B activity enhances chemical hepatocarcinogenesis through sustained c-Jun N-terminal kinase 1 activation. *Proc. Natl Acad. Sci. USA* **103,** 10544–10551 (2006).
38. Hui, L. *et al.* p38alpha suppresses normal and cancer cell proliferation by antagonizing the JNK-c-Jun pathway. *Nat. Genet.* **39,** 741–749 (2007).
39. Hui, L., Zatloukal, K., Scheuch, H., Stepniak, E. & Wagner, E. F. Proliferation of human HCC cells and chemically induced mouse liver cancers requires JNK1-dependent p21 downregulation. *J. Clin. Invest.* **118,** 3943–3953 (2008).
40. He, G. *et al.* Hepatocyte IKKbeta/NF-kappaB inhibits tumor promotion and progression by preventing oxidative stress-driven STAT3 activation. *Cancer Cell* **17,** 286–297 (2010).
41. Bennett, B. L. *et al.* SP600125, an anthrapyrazolone inhibitor of Jun N-terminal kinase. *Proc. Natl Acad. Sci. USA* **98,** 13681–13686 (2001).
42. Lan, X. *et al.* Morphine induces albuminuria by compromising podocyte integrity. *PLoS ONE* **8,** 29 (2013).
43. Nikolaou, K. *et al.* Inactivation of the deubiquitinase CYLD in hepatocytes causes apoptosis, inflammation, fibrosis, and cancer. *Cancer Cell* **21,** 738–750 (2012).
44. Shibata, S. *et al.* Adiponectin regulates cutaneous wound healing by promoting keratinocyte proliferation and migration via the ERK signaling pathway. *J. Immunol.* **189,** 3231–3241 (2012).
45. Tan, Z. *et al.* IL-17A plays a critical role in the pathogenesis of liver fibrosis through hepatic stellate cell activation. *J. Immunol.* **191,** 1835–1844 (2013).
46. Ye, L. *et al.* TRPV4 is a regulator of adipose oxidative metabolism, inflammation, and energy homeostasis. *Cell* **151,** 96–110 (2012).
47. Volker, M. *et al.* Copy number variation, chromosome rearrangement, and their association with recombination during avian evolution. *Genome Res.* **20,** 503–511 (2010).
48. Gupta, P. K. *et al.* High frequency in vivo loss of heterozygosity is primarily a consequence of mitotic recombination. *Cancer Res.* **57,** 1188–1193 (1997).
49. Burrell, R. A., McGranahan, N., Bartek, J. & Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **501,** 338–345 (2013).
50. Burrell, R. A. *et al.* Replication stress links structural and numerical cancer chromosomal instability. *Nature* **494,** 492–496 (2013).
51. Dereli-Oz, A., Versini, G. & Halazonetis, T. D. Studies of genomic copy number changes in human cancers reveal signatures of DNA replication stress. *Mol. Oncol.* **5,** 308–314 (2011).
52. Carter, S. L., Eklund, A. C., Kohane, I. S., Harris, L. N. & Szallasi, Z. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. *Nat. Genet.* **38,** 1043–1048 (2006).
53. Eltzschig, H. K. & Carmeliet, P. Hypoxia and inflammation. *New Engl. J. Med.* **364,** 656–665 (2011).
54. Coquelle, A., Toledo, F., Stern, S., Bieth, A. & Debatisse, M. A new role for hypoxia in tumor progression: induction of fragile site triggering genomic rearrangements and formation of complex DMs and HSRs. *Mol. Cell* **2,** 259–265 (1998).
55. Kumareswaran, R. *et al.* Chronic hypoxia compromises repair of DNA double-strand breaks to drive genetic instability. *J. Cell Sci.* **125,** 189–199 (2012).
56. Eferl, R. & Wagner, E. F. AP-1: a double-edged sword in tumorigenesis. *Nat. Rev. Cancer* **3,** 859–868 (2003).
57. Das, M., Garlick, D. S., Greiner, D. L. & Davis, R. J. The role of JNK in the development of hepatocellular carcinoma. *Genes Dev.* **25,** 634–645 (2011).
58. Dvorak, Z. *et al.* JNK inhibitor SP600125 is a partial agonist of human aryl hydrocarbon receptor and induces CYP1A1 and CYP1A2 genes in primary human hepatocytes. *Biochem. Pharmacol.* **75,** 580–588 (2008).
59. Marozin, S. *et al.* Posttranslational modification of vesicular stomatitis virus glycoprotein, but not JNK inhibition, is the antiviral mechanism of SP600125. *J. Virol.* **86,** 4844–4855 (2012).
60. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
61. Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22,** 568–576 (2012).

62. Ge, X. *et al.* Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. *Genomics* **86,** 127–141 (2005).
63. Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA* **101,** 6062–6067 (2004).
64. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc. Natl Acad. Sci. USA* **107,** 16910–16915 (2010).
65. Kamburov, A., Stelzl, U., Lehrach, H. & Herwig, R. The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res.* **41,** D793–D800 (2013).
66. Seglen, P. O. Preparation of isolated rat liver cells. *Methods Cell Biol.* **13,** 29–83 (1976).
67. Thoolen, B. *et al.* Proliferative and nonproliferative lesions of the rat and mouse hepatobiliary system. *Toxicol. Pathol.* **38,** 5S–81S (2010).
68. Jensen, L. J. *et al.* eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res.* **36,** D250–D254 (2008).
69. Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21,** 974–984 (2011).
70. Korbel, J. O. *et al.* PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol.* **10,** R23 (2009).

Acknowledgements

We thank the members of the Ciccarelli lab for useful discussion, Federica Pisati for histology sample preparation, Luca Giorgetti for the FISH analysis, and Sandrine Imbeaud and Jessica Zuman-Rossi for providing the alignment files of some of the samples published in ref. 3 for comparison. This project was funded by the European Union's Seventh Framework Programme (FP7/2007–2013) under grant agreement No. 259743 (MODHEP consortium). F.D.C. acknowledges the support of an Investigation Grant from the Italian Association for Cancer Research (AIRC-IG 12742) and support from the Italian Ministry of Health (Grant Giovani Ricercatori). S.G. was supported by a grant from the Italian Association for Cancer Research (AIRC-MFAG 8941).

Author contributions

F.D.C. and G.N. conceived the study; F.D.C. supervised the study; F.I. and S.S. developed all bioinformatics pipelines and analysed the data; A.C. performed the experiments; E.R. contributed the pathological inspection on mouse; P.N. assisted with the mice; L.D.A., A.S., J.F., M.A.B., E.S., R.J.T. and A.S.K. provided the human samples; A.S.K. contributed the pathological inspection on human samples; S.G. supervised the experimental work; F.D.C., F.I. and S.S. wrote the manuscript; G.N. contributed to edit the manuscript.

Additional information

Accession codes: Exome sequence data for human HCC samples have been deposited in the European Genome-phenome Archive (EGA, http://www.ebi.ac.uk/ega/) under the accession code EGAS00001000749. Whole-genome, whole-exome and targeted re-sequence data for mouse HCC samples have been deposited in GenBank/EMBL/DDBJ Sequence Read Archive (SRA) under the accession codes SRP040716, SRP040691 and SRP040712, respectively.

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Iannelli, F. *et al.* Massive gene amplification drives paediatric hepatocellular carcinoma caused by bile salt export pump deficiency. *Nat. Commun.* 5:3850 doi: 10.1038/ncomms4850 (2014).

D978–D983 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 12 November 2011*
doi:10.1093/nar/gkr952

Network of Cancer Genes (NCG 3.0): integration and analysis of genetic and network properties of cancer genes

Matteo D'Antonio, Vera Pendino, Shruti Sinha and Francesca D. Ciccarelli*

Department of Experimental Oncology, European Institute of Oncology, IFOM-IEO Campus, Via Adamello 16, 20139 Milan, Italy

Received September 13, 2011; Accepted October 12, 2011

ABSTRACT

The identification of a constantly increasing number of genes whose mutations are causally implicated in tumor initiation and progression (cancer genes) requires the development of tools to store and analyze them. The Network of Cancer Genes (NCG 3.0) collects information on 1494 cancer genes that have been found mutated in 16 different cancer types. These genes were collected from the Cancer Gene Census as well as from 18 whole exome and 11 whole-genome screenings of cancer samples. For each cancer gene, NCG 3.0 provides a summary of the gene features and the cross-reference to other databases. In addition, it describes duplicability, evolutionary origin, orthology, network properties, interaction partners, microRNA regulation and functional roles of cancer genes and of all genes that are related to them. This integrated network of information can be used to better characterize cancer genes in the context of the system in which they act. The data can also be used to identify novel candidates that share the same properties of known cancer genes and may therefore play a similar role in cancer. NCG 3.0 is freely available at http://bio.ifom-ieo-campus.it/ncg.

INTRODUCTION

The pivotal role of genomic instability in causing human cancer is an old concept that dates back to the first cytogenetic studies on cancer cells (1,2). Until very recently, however, the number of genes whose somatic mutations are causally implicated in cancer initiation and progression was very low. This was mainly due to the difficult and lengthy process of identifying genetic modifications through traditional sequencing methods. The recent development of next-generation sequencing techniques has boosted the discovery of novel cancer genes. In the last 5 years, high-throughput mutational screenings were performed on several cancer types and led to the identification of mutations in both coding and non-coding regions of the cancer genome. So far, around 30 high-throughput and whole genome sequencing experiments on cancer samples have been published, which identified around 1500 mutated genes that are potentially actively involved in cancer development (3–30). This constant delivery of new sequencing data is radically changing our understanding of cancer genetics, showing that the genomic landscape of cancer is complex and varies among and within cancer types (31). One way to reduce the complexity implies the identification of properties that are shared among cancer genes and distinguish them from the rest of human genes (32). For example, cancer genes tend to be singletons, i.e. they preserve one gene copy in the genome, and to encode hubs, i.e. proteins that engage numerous physical connections with other proteins (33). Moreover, recessive cancer genes appeared early in evolution and are involved in basic cellular processes, while oncogenes originated later and mainly act as regulators and signal transducers (34). Thus, the identification of specific properties of cancer genes may help in better rationalizing their role in cancer. While several databases collect mutational and functional information on cancer genes, no description of their systems-level properties is currently available. The Network of Cancer Genes (NCG) was originally created to collect information on duplicability and network properties of cancer genes (35).

Information on duplicability and network properties of cancer genes (35). In the current version of the database (NCG 3.0), we included 1494 cancer genes that have been identified so far, and provide an update of the human gene set and the protein interactome. We also added novel features, such as the functional description of cancer genes as well as information about their interaction with microRNAs (miRNAs). Finally, we developed a new web

*To whom correspondence should be addressed. Tel: +39 02574303053; Fax: +39 0294375990; Email: francesca.ciccarelli@ifom-ieo-campus.it

Table 1. Primary data correspond to the list of cancer genes as they were extracted from the original papers. Five genes are not present in NCG 3.0 because they could not be mapped to current version of Entrez IDs. Dominant and recessive genes refer to the definition reported in the cancer gene census (37). Ancient and recent genes refer to genes that originated between the last universal common ancestor and opisthokonts and with metazoans or later, respectively. HTMS, high-throughput mutation screening; WGS, whole genome sequencing

Cancer genes	Dominant	Recessive	Amplified	HTMS	WGS	Total
Primary data	348	98	77	699	457	1,499
Present in NCG 3	346	96	77	698	454	1,494
Duplicated (%)	58 (17.7)	10 (10.5)	23 (29.9)	102 (14.8)	64 (14.3)	230 (15.7)
Ancient (%)	184 (55.6)	63 (66.3)	53 (69.9)	377 (54.6)	273 (61.2)	837 (57.1)
Recent (%)	147 (44.4)	32 (33.7)	24 (31.1)	313 (45.4)	173 (38.8)	629 (42.9)
Hubs (%)	185 (58.6)	71 (75.5)	50 (70.4)	231 (42.6)	116 (34.6)	532 (44.7)
miRNA targets (%)	53 (15.3)	14 (14.6)	17 (22.1)	53 (7.6)	28 (6.2)	118 (7.9)
miRNA hosts (%)	12 (3.5)	6 (6.3)	2 (2.6)	28 (4.0)	17 (3.7)	55 (3.7)

interface to allow more intuitive and flexible queries of the database. Since properties of cancer genes can be also exploited to identify new candidates that may be involved in cancer (36), we present a case study that involves two paralogs, the DNA methyltransferase alpha (*DNMT3A*) and beta (*DNMT3B*). On the basis of the information that can be retrieved in NCG 3.0, we hypothesize that, upon the acquisition of somatic mutations, these genes may play an active role in cancer development.

UPDATES OF DATA SOURCES

Cancer genes

NCG 3.0 includes information on 1494 cancer genes that were extracted from 30 distinct studies (Table 1). The genes were divided into three groups, according to the experimental evidence that supported their involvement in cancer. The first group includes 498 genes from the latest release of the Cancer Gene Census (CGC, 444 genes, 31 March 2011) (37) and from the census of amplified cancer genes (77 genes) (38). Both these resources are literature-based collections of genes whose mutations and/or amplifications have been proven to be causally implicated in tumorigenesis. Among the 444 genes from CGC, 346 are defined as dominant (i.e. heterozygous mutations are sufficient to cause tumorigenesis) and 96 as recessive (i.e. homozygous mutations are required to initiate tumorigenesis) (37). The second group is formed of 698 candidate cancer genes from 18 high-throughput mutational screenings (HTMS) on 16 distinct cancer types, such as breast (15,30), colorectal (30), pancreatic (13–15), lung (10,15), renal (8), prostatic (15), bladder (12), and ovarian (15) cancers, glioblastoma (20,21), leukemia (11), lymphoma (23), sarcoma (4), myeloma (6), medulloblastoma (22), head and neck squamous cell carcinoma (3,27), and melanoma (29). These genes represent the subsets of mutated genes that were identified as bearing driver mutations, i.e. mutations that confer growth advantage and are actively involved in tumor development (31). The third group of cancer genes contains 457 genes that have been identified in eleven whole genome sequencing (WGS) screenings of cancer genomes, such as breast (26), lung (16,25), prostatic (5), liver cancer (28), glioblastoma (7), leukemia (17,19), myeloma (6) and melanoma (24).

Human gene set and orthology information

We gathered information on the human gene set from the combination of Gencode v.7.0 (39) and RefSeq v. 46 (40). Gencode is an initiative that aims at identifying and mapping all human protein coding genes (41). It has been used as the reference gene set for the Encode project, for the 1000 Genomes Project and for the design of the capture baits for whole exome sequencing (39). This gene set is therefore likely to include all genes present in current and future mutational screenings of whole cancer exomes. Starting from 20 700 genes in Gencode v.7.0 (corresponding to 84 408 protein sequences), we removed multiple isoforms that map to the same gene locus by aligning all protein sequences to the reference human genome (hg18) and retaining only the longest isoform (33). At the end of this pipeline, we were able to retrieve 19 560 unique genes. The 1140 missing genes corresponded to transcripts that spanned more than one gene. To recover them, we repeated the same pipeline with 20 750 genes in RefSeq (corresponding to 34 571 protein sequences). The union of these two gene lists led to the final dataset of 20 531 unique human genes.

To measure the duplicability of each gene, we identified all additional hits on the genome that span at least 60% of the gene length (33), and found that 4311 human genes (21% of the total) have at least one additional copy. Consistently with previous reports (33,34), we found that cancer genes duplicate significantly less than the rest of human genes (15.7% of the total, p-value $= 3.0\,10^{-7}$, Pearson's Chi-squared Test), and dominant genes are more duplicated than recessive genes (Figure 1A and Table 1).

We updated the information on orthology relationships between human and other species to eggNOG v. 2.0 (42). Orthologs were used to identify the evolutionary origin of 1,466 cancer genes (98% of the total), defined as the deepest internal node of the tree of life where an ortholog could be found (34). We confirmed that recessive cancer genes are mostly ancient genes that have orthologs in prokaryotes, plants and fungi, while a higher fraction of

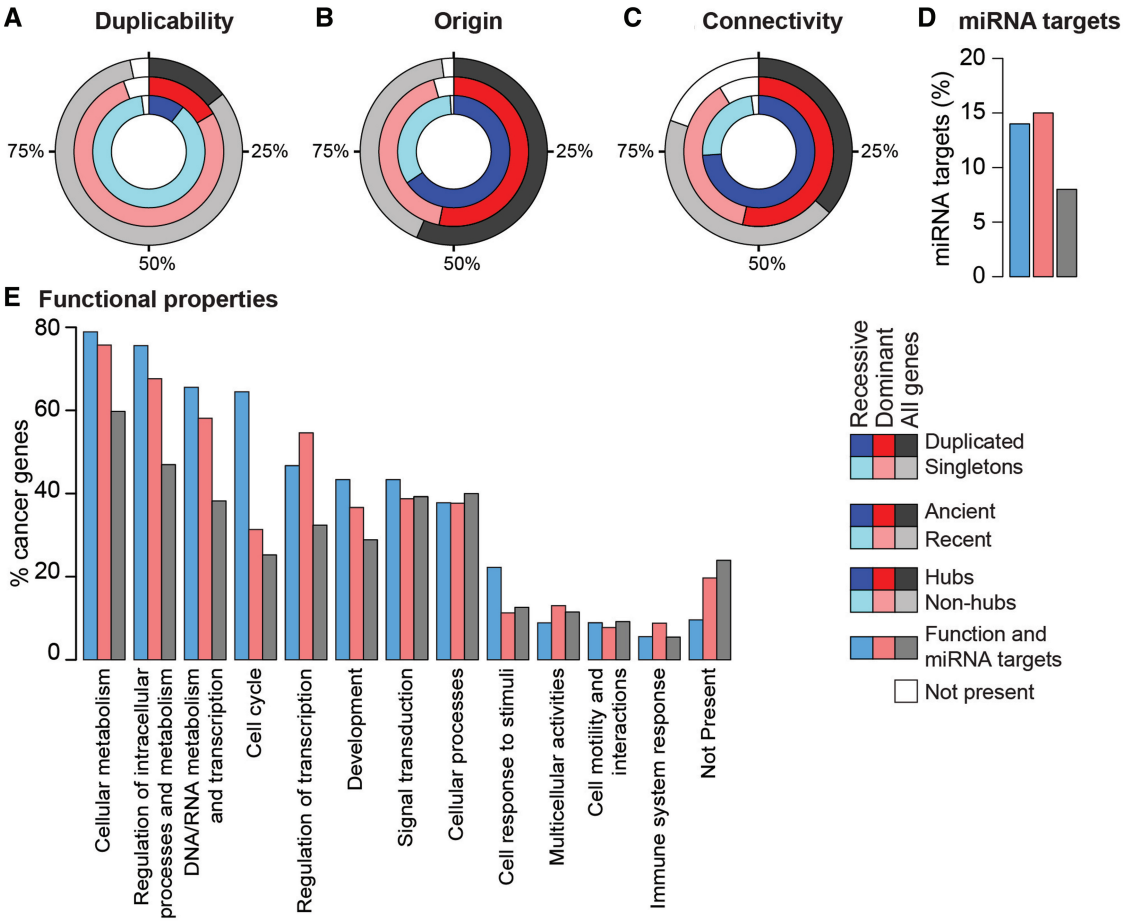


Figure 1. Properties of cancer genes. Circles represent the fraction of cancer, dominant, and recessive genes that are (**A**) singleton or duplicated, (**B**) ancient or recent, (**C**) hubs or non-hubs. Ancient genes originated between the last universal common ancestor and opisthokonts; recent genes originated with metazoans or later. (**D**) miRNA targets were derived from TarBase (50) and miRecords (51). Approximately 15% of dominant and recessive cancer genes are targets of miRNAs, compared with 8% of all cancer genes. (**E**) Cancer genes were associated to one of the 12 functional categories based on the corresponding GO terms (34).

dominant cancer genes appeared with metazoans or later (Figure 1B and Table 1).

Human interaction network

We rebuilt the human interactome by integrating the latest releases of five sources of protein–protein interactions, such as HPRD (43), BioGRID (44), IntAct (45), MINT (46) and DIP (47). Only primary interactions were used while putative interactions inferred from orthology relationships were discarded. The final dataset comprised 98 492 binary interactions between 13 531 proteins, supported by 25 915 independent publications. The resulting human interactome contains 13% more proteins and 44% more interactions than the previous version of NCG, and reports interaction data for 1189 cancer genes (79% of the total). We confirmed that cancer genes are enriched in hubs, defined as top 25% most connected proteins (at least 15 interactions) (34), and are therefore more connected than the rest of human genes (Figure 1C and Table 1).

NEW FEATURES OF NCG 3.0

Interactions between cancer gene and miRNAs

miRNAs are endogenous short nucleotide sequences that interfere with RNA transcripts in the cytoplasm, thus regulating protein synthesis (48). Alterations in miRNA expression have been reported during cancer initiation, progression and metastasis (49). In addition, miRNAs regulate well-known cancer genes, such as PTEN, NRAS, KRAS, (48). Because of this involvement in cancer, we included information about the interactions between miRNAs and their target cancer genes. Primary data were extracted from TarBase v.5.0 (50) and miRecords v.1.0 (51), which collect experimentally supported interactions between miRNAs and their targets. The non-redundant integration of the two databases returned a list of 54 miRNAs that regulate 118 cancer genes (8% of the total, Table 1). This fraction is significantly higher compared to the rest of human genes (~4%, p-value = 10^{-11}, Pearson's Chi-squared Test), and becomes even higher when only recessive or dominant

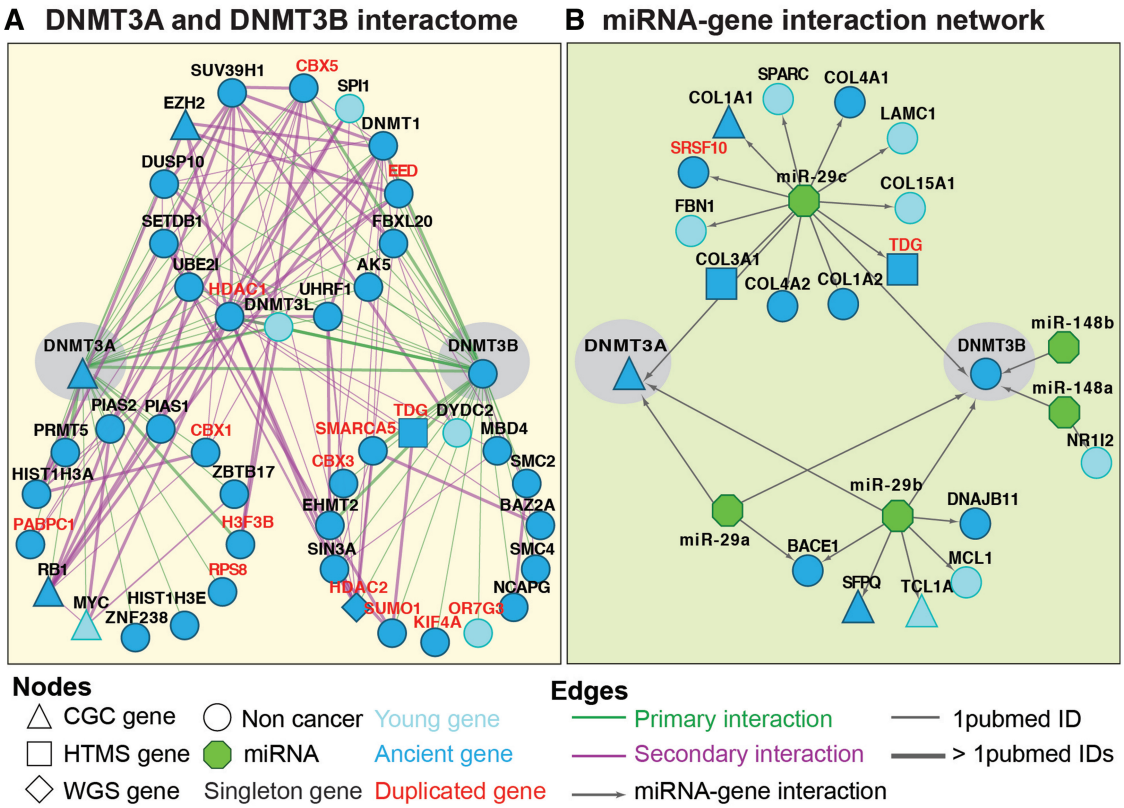


Figure 2. Functional redundancy of DNMT3A and DNMT3B. (**A**) Protein-protein interaction network of DNMT3A and DNMT3B. The two proteins share 14 interactors, which are mostly involved in the epigenetic control of chromatin and in the regulation of gene expression. Primary interactions are physical interactions of a protein directly with DNMT3A or DNMT3B, while secondary interactions are physical interactions between their primary interactors. (**B**) Interaction network between DNMT3A and DNMT3B and target miRNAs. Both genes are regulated by members of the miR-29 family, whose expression is altered in cancer and is inversely correlated with the gene expression (57). CGC, cancer gene census; HTMS, high-throughput mutational screenings; WGS, whole genome sequencing.

cancer genes are considered (~15%, Figure 1D and Table 1). This enrichment reflects at least partly the fact that cancer genes are heavily studied and therefore more information are available for them compared to other genes. In addition to miRNA targets, we also retrieved information on 55 cancer genes that host miRNAs within their genomic locus (Table 1).

Functional classification of cancer genes

To derive the gene functional classification, we first extracted 1108 cancer genes (74% of all the total) that have at least one term at levels 5 and 6 of the biological process branch of the Gene Ontology (GO) tree (52). We then grouped all these terms into 12 functional categories, in order to provide a more general description of the gene function (34). Dominant and recessive cancer genes have different functional distributions. In particular, while dominant genes are mostly involved in the regulation of transcription, recessive genes are associated with basic cellular functions such as cell cycle, cell response to stimuli, cellular and DNA/RNA metabolism (Figure 1E).

Network representation

We developed a novel web interface and added several new possibilities of querying the database, thus allowing the user to search for specific cancer genes, for lists of genes and miRNAs related to cancer, and for the presence of cancer genes in genomic regions of interest. In addition, it is now possible to retrieve the interactions between a certain miRNA and the cancer genes that are associated to it, as well as all cancer genes that are miRNA targets or that host miRNAs within their genomic locus. Finally, we adopted CytoscapeWeb (53) for the visualization of protein–protein and miRNA–target interaction networks, and for displaying the evolutionary origin of cancer genes and their orthologs in other species.

IDENTIFICATION OF CANDIDATE CANCER GENES USING NCG

In addition to providing a comprehensive description of the properties of cancer genes, NCG can be used to identify new possible candidates, based on the hypothesis that genes with systems-level properties similar to known cancer genes may also play similar roles in cancer (36). Following this idea, two almost identical paralogs, *GNAQ GNA11*, have recently been associated to the same tumor type (54). Here, we hypothesize that a similar association can be drawn for *DNMT3A* and its paralog *DNMT3B*. Both these methyltransferases cause

de novo methylation during the differentiation of hematopoietic cells and are involved in hematopoietic stem-cell renewal (55). *DNMT3A* is an already known cancer gene, part of the cancer gene census and frequently mutated in leukemia (56). Although its role in tumorigenesis is still unclear, it has been hypothesized that mutations in *DNMT3A* either induce altered gene expression or affect genome stability (56). To date there is no evidence that also *DNMT3B* is mutated in cancer, although NCG 3.0 provides several indications that this might be indeed the case. First, both *DNMT3A* and *DNMT3B* originated with eukaryotes and are associated with the same GO terms, thus indicating functional redundancy. Second, the encoded proteins DNMT3A and DNMT3B share 14 interactors, which, in turn, are connected through secondary interactions (Figure 2A). The resulting highly interconnected module involves key players in epigenetic changes of chromatin, and consequent regulation of gene expression. Third, both genes are targets of three miRNAs of the miR-29 family (Figure 2B). The downregulation of these miRNAs induces aberrant expression of both *DNMT3A* and *DNMT3B* in non-small-cell lung cancer (57) and the upregulation of both paralogs has been observed in several tumor types (58). We hypothesize that the aberrant expression can be related to the acquisition of mutations in the sequence of these genes.

ACKNOWLEDGMENTS

The authors thank all members of the Ciccarelli lab for testing NCG and providing useful suggestion to improve it.

FUNDING

Funding for open access charge: Associazione Italiana Ricerca sul Cancro (AIRC) and Fondazione Cariplo (to F.D.C.).

REFERENCES

1. von Hansemann,D.P. (1890) On primary cancer of the liver. *Berl. klin. Wschr.*, **27**, 353–356.
2. Boveri,T. (1914) Zur Frage der Entstehung Maligner Tumoren. *Jena: Gustav Fischer*, **1**, 1–64.
3. Agrawal,N., Frederick,M.J., Pickering,C.R., Bettegowda,C., Chang,K., Li,R.J., Fakhry,C., Xie,T.X., Zhang,J., Wang,J. *et al.* (2011) Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. *Science*, **333**, 1154–1157.
4. Barretina,J., Taylor,B.S., Banerji,S., Ramos,A.H., Lagos-Quintana,M., Decarolis,P.L., Shah,K., Socci,N.D., Weir,B.A., Ho,A. *et al.* (2010) Subtype-specific genomic alterations define new targets for soft-tissue sarcoma therapy. *Nat. Genet.*, **42**, 715–721.
5. Berger,M.F., Lawrence,M.S., Demichelis,F., Drier,Y., Cibulskis,K., Sivachenko,A.Y., Sboner,A., Esgueva,R., Pflueger,D., Sougnez,C. *et al.* (2011) The genomic complexity of primary human prostate cancer. *Nature*, **470**, 214–220.
6. Chapman,M.A., Lawrence,M.S., Keats,J.J., Cibulskis,K., Sougnez,C., Schinzel,A.C., Harview,C.L., Brunet,J.P., Ahmann,G.J., Adli,M. *et al.* (2011) Initial genome sequencing and analysis of multiple myeloma. *Nature*, **471**, 467–472.
7. Clark,M.J., Homer,N., O'Connor,B.D., Chen,Z., Eskin,A., Lee,H., Merriman,B. and Nelson,S.F. (2010) U87MG decoded: the genomic sequence of a cytogenetically aberrant human cancer cell line. *PLoS Genet.*, **6**, e1000832.
8. Dalgliesh,G.L., Furge,K., Greenman,C., Chen,L., Bignell,G., Butler,A., Davies,H., Edkins,S., Hardy,C., Latimer,C. *et al.* (2010) Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. *Nature*, **463**, 360–363.
9. Ding,L., Ellis,M.J., Li,S., Larson,D.E., Chen,K., Wallis,J.W., Harris,C.C., McLellan,M.D., Fulton,R.S., Fulton,L.L. *et al.* (2010) Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature*, **464**, 999–1005.
10. Ding,L., Getz,G., Wheeler,D.A., Mardis,E.R., McLellan,M.D., Cibulskis,K., Sougnez,C., Greulich,H., Muzny,D.M., Morgan,M.B. *et al.* (2008) Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, **455**, 1069–1075.
11. Greif,P.A., Eck,S.H., Konstandin,N.P., Benet-Pages,A., Ksienzyk,B., Dufour,A., Vetter,A.T., Popp,H.D., Lorenz-Depiereux,B., Meitinger,T. *et al.* (2011) Identification of recurring tumor-specific somatic mutations in acute myeloid leukemia by transcriptome sequencing. *Leukemia*, **25**, 821–827.
12. Gui,Y., Guo,G., Huang,Y., Hu,X., Tang,A., Gao,S., Wu,R., Chen,C., Li,X., Zhou,L. *et al.* (2011) Frequent mutations of chromatin remodeling genes in transitional cell carcinoma of the bladder. *Nat. Genet.*, **43**, 875–878.
13. Jiao,Y., Shi,C., Edil,B.H., de Wilde,R.F., Klimstra,D.S., Maitra,A., Schulick,R.D., Tang,L.H., Wolfgang,C.L., Choti,M.A. *et al.* (2011) DAXX/ATRX, MEN1, and mTOR pathway genes are frequently altered in pancreatic neuroendocrine tumors. *Science*, **331**, 1199–1203.
14. Jones,S., Zhang,X., Parsons,D.W., Lin,J.C., Leary,R.J., Angenendt,P., Mankoo,P., Carter,H., Kamiyama,H., Jimeno,A. *et al.* (2008) Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science*, **321**, 1801–1806.
15. Kan,Z., Jaiswal,B.S., Stinson,J., Janakiraman,V., Bhatt,D., Stern,H.M., Yue,P., Haverty,P.M., Bourgon,R., Zheng,J. *et al.* (2010) Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature*, **466**, 869–873.
16. Lee,W., Jiang,Z., Liu,J., Haverty,P.M., Guan,Y., Stinson,J., Yue,P., Zhang,Y., Pant,K.P., Bhatt,D. *et al.* (2010) The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature*, **465**, 473–477.
17. Ley,T.J., Mardis,E.R., Ding,L., Fulton,B., McLellan,M.D., Chen,K., Dooling,D., Dunford-Shore,B.H., McGrath,S., Hickenbotham,M. *et al.* (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*, **456**, 66–72.
18. Li,M., Zhao,H., Zhang,X., Wood,L.D., Anders,R.A., Choti,M.A., Pawlik,T.M., Daniel,H.D., Kannangai,R., Offerhaus,G.J. *et al.* (2011) Inactivating mutations of the chromatin remodeling gene ARID2 in hepatocellular carcinoma. *Nat. Genet.*, **43**, 828–829.
19. Mardis,E.R., Ding,L., Dooling,D.J., Larson,D.E., McLellan,M.D., Chen,K., Koboldt,D.C., Fulton,R.S., Delehaunty,K.D., McGrath,S.D. *et al.* (2009) Recurring mutations found by sequencing an acute myeloid leukemia genome. *N. Engl. J. Med.*, **361**, 1058–1066.
20. McLendon,R., Friedman,A., Bigner,D., Van,M.E.G., Brat,D., Mastrogianakis,G., Olson,J., Mikkelsen,T., Lehman,N., Aldape,K. *et al.* (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.
21. Parsons,D.W., Jones,S., Zhang,X., Lin,J.C., Leary,R.J., Angenendt,P., Mankoo,P., Carter,H., Siu,I.M., Gallia,G.L. *et al.* (2008) An integrated genomic analysis of human glioblastoma multiforme. *Science*, **321**, 1807–1812.
22. Parsons,D.W., Li,M., Zhang,X., Jones,S., Leary,R.J., Lin,J.C., Boca,S.M., Carter,H., Samayoa,J., Bettegowda,C. *et al.* (2010) The genetic landscape of the childhood cancer medulloblastoma. *Science*, **331**, 435–439.
23. Pasqualucci,L., Trifonov,V., Fabbri,G., Ma,J., Rossi,D., Chiarenza,A., Wells,V.A., Grunn,A., Messina,M., Elliot,O. *et al.* (2011) Analysis of the coding genome of diffuse large B-cell lymphoma. *Nat. Genet.*

24. Pleasance,E.D., Cheetham,R.K., Stephens,P.J., McBride,D.J., Humphray,S.J., Greenman,C.D., Varela,I., Lin,M.L., Ordonez,G.R., Bignell,G.R. *et al.* (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, **463**, 191–196.
25. Pleasance,E.D., Stephens,P.J., O'Meara,S., McBride,D.J., Meynert,A., Jones,D., Lin,M.L., Beare,D., Lau,K.W., Greenman,C. *et al.* (2010) A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature*, **463**, 184–190.
26. Shah,S.P., Morin,R.D., Khattra,J., Prentice,L., Pugh,T., Burleigh,A., Delaney,A., Gelmon,K., Guliany,R., Senz,J. *et al.* (2009) Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature*, **461**, 809–813.
27. Stransky,N., Egloff,A.M., Tward,A.D., Kostic,A.D., Cibulskis,K., Sivachenko,A., Kryukov,G.V., Lawrence,M., Sougnez,C., McKenna,A. *et al.* (2011) The mutational landscape of head and neck squamous cell carcinoma. *Science*, **333**, 1157–1160.
28. Totoki,Y., Tatsuno,K., Yamamoto,S., Arai,Y., Hosoda,F., Ishikawa,S., Tsutsumi,S., Sonoda,K., Totsuka,H., Shirakihara,T. *et al.* (2011) High resolution characterization of a hepatocellular carcinoma genome. *Nat. Genet.*, **43**, 464–469.
29. Wei,X., Walia,V., Lin,J.C., Teer,J.K., Prickett,T.D., Gartner,J., Davis,S., Stemke-Hale,K., Davies,M.A., Gershenwald,J.E. *et al.* (2011) Exome sequencing identifies GRIN2A as frequently mutated in melanoma. *Nat. Genet.*, **43**, 442–446.
30. Wood,L.D., Parsons,D.W., Jones,S., Lin,J., Sjoblom,T., Leary,R.J., Shen,D., Boca,S.M., Barber,T., Ptak,J. *et al.* (2007) The genomic landscapes of human breast and colorectal cancers. *Science*, **318**, 1108–1113.
31. Stratton,M.R., Campbell,P.J. and Futreal,P.A. (2009) The cancer genome. *Nature*, **458**, 719–724.
32. Ciccarelli,F.D. (2010) The (r)evolution of cancer genetics. *BMC Biol.*, **8**, 74.
33. Rambaldi,D., Giorgi,F.M., Capuani,F., Ciliberto,A. and Ciccarelli,F.D. (2008) Low duplicability and network fragility of cancer genes. *Trends Genet.*, **24**, 427–430.
34. D'Antonio,M. and Ciccarelli,F.D. (2011) Modification of gene duplicability during the evolution of protein interaction network. *PLoS Comput. Biol.*, **7**, e1002029.
35. Syed,A.S., D'Antonio,M. and Ciccarelli,F.D. (2010) Network of cancer genes: a web resource to analyze duplicability, orthology and network properties of cancer genes. *Nucleic Acids Res.*, **38**, D670–D675.
36. Anelli,V., Santoriello,C., Distel,M., Koster,R.W., Ciccarelli,F.D. and Mione,M. (2009) Global repression of cancer gene expression in a zebrafish model of melanoma is linked to epigenetic regulation. *Zebrafish*, **6**, 417–424.
37. Futreal,P.A., Coin,L., Marshall,M., Down,T., Hubbard,T., Wooster,R., Rahman,N. and Stratton,M.R. (2004) A census of human cancer genes. *Nat. Rev. Cancer*, **4**, 177–183.
38. Santarius,T., Shipley,J., Brewer,D., Stratton,M.R. and Cooper,C.S. (2010) A census of amplified and overexpressed human cancer genes. *Nat. Rev. Cancer*, **10**, 59–64.
39. Coffey,A.J., Kokocinski,F., Calafato,M.S., Scott,C.E., Palta,P., Drury,E., Joyce,C.J., Leproust,E.M., Harrow,J., Hunt,S. *et al.* (2011) The GENCODE exome: sequencing the complete human exome. *Eur. J. Hum. Genet.*, **19**, 827–831.
40. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
41. Harrow,J., Denoeud,F., Frankish,A., Reymond,A., Chen,C.K., Chrast,J., Lagarde,J., Gilbert,J.G., Storey,R., Swarbreck,D. *et al.* (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biol.*, **7(Suppl. 1)**, S41–S49.
42. Muller,J., Szklarczyk,D., Julien,P., Letunic,I., Roth,A., Kuhn,M., Powell,S., von Mering,C., Doerks,T., Jensen,L.J. *et al.* (2010) eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res.*, **38**, D190–D195.
43. Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A. *et al.* (2009) Human protein reference database–2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
44. Stark,C., Breitkreutz,B.J., Chatr-Aryamontri,A., Boucher,L., Oughtred,R., Livstone,M.S., Nixon,J., Van Auken,K., Wang,X., Shi,X. *et al.* (2010) The BioGRID interaction database: 2011 update. *Nucleic Acids Res.*, **39**, D698–D704.
45. Aranda,B., Achuthan,P., Alam-Faruque,Y., Armean,I., Bridge,A., Derow,C., Feuermann,M., Ghanbarian,A.T., Kerrien,S., Khadake,J. *et al.* (2009) The IntAct molecular interaction database in 2010. *Nucleic Acids Res.*, **38**, D525–D531.
46. Ceol,A., Chatr Aryamontri,A., Licata,L., Peluso,D., Briganti,L., Perfetto,L., Castagnoli,L. and Cesareni,G. (2009) MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res.*, **38**, D532–D539.
47. Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.
48. Inui,M., Martello,G. and Piccolo,S. (2010) MicroRNA control of signal transduction. *Nat. Rev. Mol. Cell. Biol.*, **11**, 252–263.
49. Croce,C.M. (2009) Causes and consequences of microRNA dysregulation in cancer. *Nat. Rev. Genet.*, **10**, 704–714.
50. Papadopoulos,G.L., Reczko,M., Simossis,V.A., Sethupathy,P. and Hatzigeorgiou,A.G. (2009) The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res.*, **37**, D155–D158.
51. Xiao,F., Zuo,Z., Cai,G., Kang,S., Gao,X. and Li,T. (2009) miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res.*, **37**, D105–D110.
52. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
53. Lopes,C.T., Franz,M., Kazi,F., Donaldson,S.L., Morris,Q. and Bader,G.D. (2010) Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, **26**, 2347–2348.
54. Van Raamsdonk,C.D., Griewank,K.G., Crosby,M.B., Garrido,M.C., Vemula,S., Wiesner,T., Obenauf,A.C., Wackernagel,W., Green,G., Bouvier,N. *et al.* (2010) Mutations in GNA11 in uveal melanoma. *N. Engl. J. Med.*, **363**, 2191–2199.
55. Tadokoro,Y., Ema,H., Okano,M., Li,E. and Nakauchi,H. (2007) De novo DNA methyltransferase is essential for self-renewal, but not for differentiation, in hematopoietic stem cells. *J. Exp. Med.*, **204**, 715–722.
56. Ley,T.J., Ding,L., Walter,M.J., McLellan,M.D., Lamprecht,T., Larson,D.E., Kandoth,C., Payton,J.E., Baty,J., Welch,J. *et al.* (2010) DNMT3A mutations in acute myeloid leukemia. *N. Engl. J. Med.*, **363**, 2424–2433.
57. Fabbri,M., Garzon,R., Cimmino,A., Liu,Z., Zanesi,N., Callegari,E., Liu,S., Alder,H., Costinean,S., Fernandez-Cymering,C. *et al.* (2007) MicroRNA-29 family reverts aberrant methylation in lung cancer by targeting DNA methyltransferases 3A and 3B. *Proc. Natl Acad. Sci. USA*, **104**, 15805–15810.
58. Li,B., Wang,B., Niu,L.J., Jiang,L. and Qiu,C.C. (2011) Hypermethylation of multiple tumor-related genes associated with DNMT3b up-regulation served as a biomarker for early diagnosis of esophageal squamous cell carcinoma. *Epigenetics*, **6**, 307–316.

GeneCNV: detection of gene copy number variations and loss of heterozygosity from whole exome sequencing data

Shruti Sinha[1], Fabio Iannelli[2] , Gennaro Gambardella[1], Matteo Cereda[1], Francesca D. Ciccarelli[1§]

[1] Division of Cancer Studies, King's College London, London SE1 1UL, UK

[2] Department of Experimental Oncology, European Institute of Oncology (IEO), IFOM-IEO Campus, Via Adamello 16, 20139 Milan, Italy

[§]Corresponding author

Email addresses:

SS: shruti.sinha@kcl.ac.uk

FI: fabio.iannelli@ieo.eu

GG: gennaro.1.gambardella@kcl.ac.uk

MC: matteo.cereda@kcl.ac.uk

FDC: francesca.ciccarelli@kcl.ac.uk

Abstract

Background: The detection of copy number variations (CNVs) and copy neutral loss of heterozygosity (CN-LOH) events from exome sequencing data is challenging because of the small size and sparse distribution of exons in the genome. As a result, available methods often fail to detect alterations of single genes. Here we present GeneCNV, a novel method suited at identifying CNVs and CN-LOHs at the gene level.

Methods: GeneCNV concatenates targeted exons to rebuild full-length genes, measures the normalized coverage of each gene independently, and finally derives sample-specific thresholds to identify amplified, deleted, and CN-LOH genes as compared to a reference sequence.

Results and Conclusions: GeneCNV shows the best performance among exome-based methods and the highest concordance with SNP array results. It also provides a graphical summary that allows a comparative and analytical assessment of results.

Keywords

Copy number variations, cancer genomics, whole exome sequencing

Background

Copy number variations (CNVs) are genomic alterations that lead to quantitative changes of the genomic content and account for ~12% of human genomic variability [1]. Several inherited CNVs have been associated with the onset of genetic disorders, including glomerulonephritis, autism, and rheumatoid arthritis [2-5]. Somatically acquired CNVs and copy neutral loss of heterozygosity (CN-LOH) events frequently occur in cancer [6, 7], where they can drive tumor progression [8-12]. Because of their primary roles in genome evolution and in disease onset, several technologies have been developed to identify CNVs, including single nucleotide polymorphism (SNP) array, array comparative genomic hybridizations (aCGH), and, more recently, next generation sequencing (NGS). Although whole genome sequencing (WGS) provides the most comprehensive CNV profile, whole exome sequencing (WES) remains the most widely used NGS approach because it is still time and cost effective. In addition, WES gives insights into the genomic alterations of protein coding genes, which are often the most interesting to follow up. CNV detection from WES data is however challenging due to the different size and sequence composition of exons, which result in non-uniform sequence coverage [13-15]. In addition, since exons are located at variable distances from each other, the detection of CNVs at breakpoint resolution is challenging [16, 17]. Several methods have been developed in recent years to circumvent these issues, including ExomeCNV [18], VarScan2 [19], ControlFreeC [20], ADTEx [21], and EXCAVATOR [22]. Although using different technical solutions, the majority of these methods identify CNV regions between two or more samples after segmenting the genome sequence into portions. Segmentation works by merging adjacent genomic regions into longer segments in the attempt to minimize coverage variability within the individual segments while maximizing it

between them. Segmentation has been widely applied to aCGH [23, 24] and WGS [25, 26] data, where information is available for long and contiguous regions. Although it has been adapted also to detect CNVs from exome data, the scattered nature of the data and the high variability in exon coverage reduce its efficacy in this context.

Here we present GeneCNV, a novel method that detects CNVs and CN-LOHs without relying on segmentation. GeneCNV starts by rebuilding full-length targeted genes. It then calculates the coverage of each gene individually and normalizes it to reduce the variability within and between samples. Finally, it uses the frequency of germline heterozygous mutations to identify sample-specific thresholds for calling amplified, deleted, and CN-LOH genes at the sample level. When compared to other exome-based methods, GeneCNV shows the highest concordance with SNP array-derived CNVs and provides the best trade-off between sensitivity and specificity.

Methods

The input data for GeneCNV are WES data from test and reference samples. Although test and reference may be of any kind, hereon we refer to them as tumor and normal exomes for the sake of clarity.

Gene coverage calculation and normalization

The coverage of each targeted exon is calculated using CoverageBed from BEDTool [27] as:

$$ExonCoverage_e = \sum_{d=0}^{\max(depth)} d \times b_d$$

where d is the depth of coverage and b_d is the number of bases at depth of coverage d in exon e. Exons of each gene are merged using the targeted gene annotation file (*e.g.* the Agilent SureSelect bed files or equivalent) to rebuild the full-length gene. Gene coverage is measured as the cumulative coverage of all exons divided by the gene length:

$$GeneCoverage_g = \frac{\sum_{e=1}^{no.\,of\,exons} ExonCoverage_e}{\sum_{e=1}^{no.\,of\,exons} ExonLength_e}$$

where e is the exon and g is the gene.

To minimize the variability of capture and sequencing efficiency within the exome, median scaling normalization [28, 29] is applied:

$$GeneCoverage_g' = \frac{\sqrt{GeneCoverage_g}}{median(\sqrt{GeneCoverage_g})_s}$$

where g is the gene and s is the sample.

To correct for gene coverage variations between tumor and matched normal exomes, quantile normalization [30] is applied. Genes in the tumor and in the matched normal exomes are ranked according to their normalized gene coverage values. The coverage of genes occupying equivalent positions in the two ranked lists is then reassigned as the average gene coverage between the two values ($GeneCoverage_g''$). For each gene the $\log_2$ratio between the gene coverage in the tumor and in the matched normal exome ($L2R_{GC}$) is finally calculated as:

$$L2R_{GC} = Log_2\left(\frac{GeneCoverage''_{g,tumor}}{GeneCoverage''_{g,normal}}\right)$$

Identification of diploid copy number regions and sample specific $L2R_{GC}$ thresholds

In order to identify sample-specific thresholds of $L2R_{GC}$ for calling amplified and deleted genes, GeneCNV relies on the deviation from the expected frequency of heterozygous SNPs (frequency = 50%) in cases of allelic imbalance due to CNVs. Heterozygous SNPs are first identified as germline mutations with frequency within 40-60% interval in the normal sample. Tumor exome is then divided into non-overlapping regions each containing 100 such heterozygous SNPs. Regions hosting ≥80% of heterozygous SNPs within 40-60% frequency are considered to maintain the allelic balance. Since high $L2R_{GC}$ values often indicate balanced amplifications of both the alleles, regions with $L2R_{GC}$ values in the lower tail (<10%) of the distribution of $L2R_{GC}$ values of all allelic balanced regions are considered as diploid. The thresholds for calling amplified and deleted genes are finally calculated from the distribution of $L2R_{GC}$ in the diploid regions as:

$$L2R_{GCA} = \overline{(L2R_{GC})}_{10\%} + 1SD_{(L2R_{GC})_{10\%}}$$

$$L2R_{GCD} = \overline{(L2R_{GC})}_{10\%} - 1SD_{(L2R_{GC})_{10\%}}$$

where $1SD$ represents one standard deviation from the $L2R_{GC}$ value.

The optimal numbers of SNPs to divide the exome (100 SNPs), the $L2R_{GC}$ value to define diploid regions (10% of the distribution) and the number of standard deviations ($1SD_{L2R_{GC}}$) have been empirically estimated and are given as the default parameters. They can be however modified by the user**.**

Identification of altered genes from SNP array

SNP arrays of 28 tumors and matched normal samples were downloaded from the Gene Expression Omnibus (GEO, GSE31174) [31, 32] and from the European

Genome-phenome Archive (EGA, EGAS00001000749) [33, 34] databases. All data were analyzed using ASCAT (version 2.1) [35] with default parameters (segment length = 25; homozygous probes with 0.3>B allele frequency>0.7 in the normal were masked). High confidence somatic CNVs and CN-LOHs were identified as genomic segments with an aberration reliability score higher than the ASCAT average reliability score in each tumor sample. The genomic coordinates of CNV and CN-LOH regions were intersected with those of the human gene sets of the Agilent SureSelect Human All Exon 50Mb kit (20,965 genes) or 38Mb kit (19,104 genes) depending on the kit used to capture the exome in the corresponding sample. A gene was considered as altered if ≥80% of its length was contained in a CNV or CN-LOH region.

Identification of altered genes from exome sequencing data

Exome sequencing data for the 28 tumors and matched normal samples were downloaded from Sequence Read Archive (SRA, DRA000433) [32, 36] and from EGA (EGAS00001000749) [34] databases. Sequencing reads from each tumor and matched normal were mapped to the human genome (GRCh37/hg19) using Novoalign [37] or BWA [38]. At most three mismatches per read were allowed and duplicated reads were removed using rmdup of SAMtools [39]. Only reads uniquely mapping within 75-100 bp of the targeted regions were considered on target and retained for further analysis. CNVs from exome sequencing data were identified in each of the 28 tumor exomes using GeneCNV, ExomeCNV (version 1.4), VarScan 2 (version 2.3.6) and EXCAVATOR (version 2.2). All methods were run using default parameters (ExomeCNV: minimum sensitivity and specificity = 0.9999 and optimizing for specificity; VarScan 2: minimum coverage = 8, minimum size = 10 bases, log ratio

threshold = 0.25 for both the lower and upper bounds; EXCAVATOR: mode = somatic). Since ExomeCNV, VarScan 2 and EXCAVATOR identify CNV regions, to detect deleted and amplified genes in each tumor exome similar analysis was done as previously described for the SNP arrays. CN-LOH genes could only be identified using ExomeCNV and GeneCNV.

Performance comparison among exome-based methods

To compare the performance of the four exome-based methods, CNV and CN-LOH genes from the SNP arrays occurring in the tumor samples were considered as the true alterations. Sensitivity, specificity, accuracy and Jaccard index was then calculated for each exome-based method as compared to SNP array results. True positives were defined as genes with the same alterations as detected in the SNP arrays and true negatives were considered as unaltered genes in both. Sensitivity was calculated as the number of true positives over the total number of altered genes in the SNP arrays. Specificity was estimated as the number of true negatives over the total number of unaltered genes in the SNP arrays. Accuracy was measured as the number of correct calls (sum of true positives and true negatives) over the total number of targeted genes. The concordance between the results from exome-based methods and SNP arrays was measured using the Jaccard index as the number of true positives over the sum of altered genes detected by the exome-based method and in SNP arrays. B-allele frequency (BAF) and log R ratio (LRR) were used to identify clonal events. Namely, tumor alterations hosting SNPs with 0.4>BAF >0.6 or with -0.25>LRR>0.25 were considered as clonal events [40, 41].

Results and Discussion

GeneCNV rationale and workflow

To minimize coverage variability due to exon length, composition, and discrete distribution along the genome, GeneCNV adopts a novel strategy that differs from those of other exome-based methods (Figure 1). First, it merges all targeted exons of each gene to rebuild a contiguous region that spans the entire gene length. This is then used to calculate the average gene coverage, thus allowing uniformly captured and sequenced exons to compensate for the non-uniform coverage of other exons within the same gene (Figure 1A). Second, it normalizes the gene coverage between genes within the same exome using global median normalization and across tumor and matched normal exomes using quantile normalization (Figure 1B). The two normalizations minimize the coverage variability due to sequence composition, DNA quality, library preparation protocols, and sequencing settings and performance. Using the normalized gene coverage in the tumor and normal exomes, GeneCNV calculates the gene coverage $\log_2$ratio ($L2R_{GC}$) between tumor and matched normal for each targeted gene (Figure 1C). In principle, $L2R_{GC}$ values around zero indicate genes with the same copy number in the tumor and in the normal counterpart, while $L2R_{GC}$ values higher than 0.6 (corresponding to $\geq$1.5 fold change) or lower than -1 (corresponding to $\leq$0.5 fold change) indicate tumor-specific gene amplifications and deletions, respectively. Such fixed thresholds, however, can be used only in genomes where no widespread chromosomal rearrangements occur. Since most cancer genomes acquire significant chromosomal instability, they often show substantially modified $L2R_{GC}$ spectra. For this reason, GeneCNV does not apply fixed $L2R_{GC}$ thresholds but derives them for each tumor sample. To this aim, it identifies regions of the tumor genome where the frequency of heterozygous SNPs is around 50%, thus

indicating that their allelic balance is maintained (see Methods and Figure 1D). Within these allelic balanced regions, GeneCNV identifies diploid regions as the ones with $L2R_{GC}$ values in the lower tail of the $L2R_{GC}$ distribution (Figure 1E), because regions of allelic balance but with high $L2R_{GC}$ values correspond to tumor-specific amplifications of both alleles. GeneCNV then uses the distribution of $L2R_{GC}$ values in diploid regions to identify sample-specific $L2R_{GC}$ thresholds for amplification ($L2R_{GCA}$) and deletion ($L2R_{GCD}$, Figure 1F). Genes in region of allelic imbalance and with $L2R_{GC}$ values within the thresholds are considered as undergoing CN-LOH. All the other genes are regarded as maintaining a two-copy status (Figure 1G).

Comparison of GeneCNV with other methods

We assessed the performance of GeneCNV as compared to three other widely used exome-based methods (ExomeCNV, VarScan 2, and EXCAVATOR). As a test dataset, we used tumor and matched normal WES data from 22 myelodysplasias [32] and six pediatric hepatocellular carcinomas [34] (Additional file 1). Tumor-specific amplified, deleted and CN-LOH genes detected from the SNP arrays on the same samples were used as the gold standard for comparison (Additional file 2). For each of the four exome-based methods, we measured sensitivity, specificity, accuracy, and the Jaccard index, which estimates the concordance with the SNP array results. Overall, GeneCNV showed the highest sensitivity (51%), accuracy (78%), and Jaccard index (36%) as compared to the other three methods (Figure 2A and Additional File 3). In addition, although it had the second best specificity (88%), it showed the best trade-off between specificity and sensitivity (Figure 2B). All four exome-based methods in general, and GeneCNV in particular, were more sensitive in detecting amplified genes (Figure 2C) than deleted genes (Figure 2D). GeneCNV

again showed the highest concordance with the SNP array results in detecting deletions (Figure 2D), thus suggesting that the higher sensitivity of the other methods was due to an overall overestimation of deletions. Of the four exome-based method, only GeneCNV and ExomeCNV can detect CN-LOHs, and GeneCNV showed the best performance in all comparisons (Figure 2E). Overall, exome-based methods had poor concordance with somatic copy number events called by SNP arrays (Figure 2A). In order to understand the reasons for this, we assessed the performances in each tumor exome individually and noticed that exome-based methods consistently failed to detect any CNVs in some tumors (Additional file 4). Since the SNP call rate of some of them (e.g. MDS-15, tAML-07 and tAML-02, Additional file 1) was low, this suggested the occurrence of possible false positives in the SNP arrays. Moreover, tumor somatic alterations are usually a mixture of clonal and subclonal events, depending on when they occur during cancer growth. Sensitivity of exome-based method is lower in detecting subclonal CNVs, because the differences in coverage between tumor and normal samples are not high enough to identify changes in copy number. This in principle leads to false negative calls from exome-based methods. To understand whether this was the case, we defined clonal events on the basis of the B allele frequency and log R ratio of heterozygous SNPs in the 28 tumor exomes and re-assessed the performances of exome-based methods only on clonal events (see Methods and Additional file 5). We noticed substantial increase in sensitivity, while specificity remained unchanged (Figure 2F and Additional file 6).

GeneCNV graphical output

In addition to the list of amplified, deleted and CN-LOH genes, GeneCNV also generates a graphical output that provides an analytical report of the results (Figure 3).

This report includes the variation of gene coverage in the tumor and matched normal before and after normalization, which can be used to evaluate the coverage differences between the two exomes (Figure 3A). It also shows the $L2R_{GC}$ spectrum of all targeted genes in the tumor exome, with amplified, deleted, and CN-LOH genes depicted in green, red, and yellow, respectively (Figure 3B). It then provides a quantification of tumor-altered genes and the cumulative density map of their distribution along the chromosomes (Figure 3C). Finally, it summarizes all results in a circos plot where altered genes that play known [42] or candidate [43] roles in cancer are shown (Figure 3D). GeneCNV can also be used to compare CNV profiles across multiple samples, which is useful when a cohort of patients is screened. In this case, the report summarizes the percentage of total altered genes in each tumor exome (Figure 3E), highlights frequently altered genes across all samples (Figure 3F) and provides the corresponding circos plot (Figure 3G). This is particularly useful to pinpoint possible cancer driver events.

Conclusions

In this study, we present GeneCNV, a novel method that detects gene-specific CNVs and CN-LOHs from WES data, based on normalized gene coverage and sample-specific thresholds. GeneCNV does not rely on segmentation of the genome to detect genomic alterations and this leads to better performances as compared to other exome-based methods. GeneCNV is suited to analyse tumor-normal exome pairs as well as larger cohorts of samples.

Authors' contributions

SS designed and developed the software, analyzed the data and compared the performance with other methods; FI supervised the method and tested the software; GG and MC contributed to the software development; FDC designed and supervised the study. FDC and SS wrote the manuscript. All authors commented on the manuscript.

Acknowledgements

The authors thank the members of the Ciccarelli lab for useful discussion. This project was funded by the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 259743 (MODHEP consortium). SS is student of the European School of Molecular Medicine (SEMM) enrolled in University of Milan and is supported by the MODHEP consortium and the Lifelong Learning Programme-Erasmus Placement of the University of Milan.

References

1. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, Gonzalez JR, Gratacos M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, et al: **Global variation in copy number in the human genome.** *Nature* 2006, **444:**444-454.
2. Aitman TJ, Dong R, Vyse TJ, Norsworthy PJ, Johnson MD, Smith J, Mangion J, Roberton-Lowe C, Marshall AJ, Petretto E, Hodges MD, Bhangal G, Patel SG, Sheehan-Rooney K, Duda M, Cook PR, Evans DJ, Domin J, Flint J, Boyle JJ, Pusey CD, Cook HT: **Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans.** *Nature* 2006, **439:**851-855.
3. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, Leotta A, Pai D, Zhang R, Lee YH, Hicks J, Spence SJ, Lee AT, Puura K, Lehtimaki T, Ledbetter D, Gregersen PK, Bregman J, Sutcliffe JS, Jobanputra V, Chung W, Warburton D, King MC, Skuse D, Geschwind DH, Gilliam TC, et al: **Strong association of de novo copy number mutations with autism.** *Science* 2007, **316:**445-449.
4. Sharp AJ, Hansen S, Selzer RR, Cheng Z, Regan R, Hurst JA, Stewart H, Price SM, Blair E, Hennekam RC, Fitzpatrick CA, Segraves R, Richmond TA, Guiver C, Albertson DG, Pinkel D, Eis PS, Schwartz S, Knight SJ, Eichler EE: **Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome.** *Nat Genet* 2006, **38:**1038-1042.
5. McKinney C, Merriman ME, Chapman PT, Gow PJ, Harrison AA, Highton J, Jones PB, McLean L, O'Donnell JL, Pokorny V, Spellerberg M, Stamp LK, Willis J, Steer S, Merriman TR: **Evidence for an influence of chemokine ligand 3-like 1 (CCL3L1) gene copy number on susceptibility to rheumatoid arthritis.** *Ann Rheum Dis* 2008, **67:**409-413.
6. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C: **Emerging landscape of oncogenic signatures across human cancers.** *Nat Genet* 2013, **45:**1127-1133.
7. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhang CZ, Wala J, Mermel CH, Sougnez C, Gabriel SB, Hernandez B, Shen H, Laird PW, Getz G, Meyerson M, Beroukhim R: **Pan-cancer patterns of somatic copy number alteration.** *Nat Genet* 2013, **45:**1134-1140.
8. Trotman LC, Niki M, Dotan ZA, Koutcher JA, Di Cristofano A, Xiao A, Khoo AS, Roy-Burman P, Greenberg NM, Van Dyke T, Cordon-Cardo C, Pandolfi PP: **Pten dose dictates cancer progression in the prostate.** *PLoS Biol* 2003, **1:**E59.
9. Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL: **Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene.** *Science* 1987, **235:**177-182.
10. Diskin SJ, Hou C, Glessner JT, Attiyeh EF, Laudenslager M, Bosse K, Cole K, Mosse YP, Wood A, Lynch JE, Pecor K, Diamond M, Winter C, Wang K,

Kim C, Geiger EA, McGrady PW, Blakemore AI, London WB, Shaikh TH, Bradfield J, Grant SF, Li H, Devoto M, Rappaport ER, Hakonarson H, Maris JM: **Copy number variation at 1q21.1 associated with neuroblastoma.** *Nature* 2009, **459:**987-991.

11. Zhuang Z, Park WS, Pack S, Schmidt L, Vortmeyer AO, Pak E, Pham T, Weil RJ, Candidus S, Lubensky IA, Linehan WM, Zbar B, Weirich G: **Trisomy 7-harbouring non-random duplication of the mutant MET allele in hereditary papillary renal carcinomas.** *Nat Genet* 1998, **20:**66-69.
12. Yoshimoto M, Cutz JC, Nuin PA, Joshua AM, Bayani J, Evans AJ, Zielenska M, Squire JA: **Interphase FISH analysis of PTEN in histologic sections shows genomic deletions in 68% of primary prostate cancer and 23% of high-grade prostatic intra-epithelial neoplasias.** *Cancer Genet Cytogenet* 2006, **169:**128-137.
13. Magi A, Tattini L, Pippucci T, Torricelli F, Benelli M: **Read count approach for DNA copy number variants detection.** *Bioinformatics* 2012, **28:**470-478.
14. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP: **Sequencing depth and coverage: key considerations in genomic analyses.** *Nat Rev Genet* 2014, **15:**121-132.
15. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ, McCombie WR: **Genome-wide in situ exon capture for selective resequencing.** *Nat Genet* 2007, **39:**1522-1527.
16. Liu B, Morrison CD, Johnson CS, Trump DL, Qin M, Conroy JC, Wang J, Liu S: **Computational methods for detecting copy number variations in cancer genome using next generation sequencing: principles and challenges.** *Oncotarget* 2013, **4:**1868-1881.
17. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z: **Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives.** *BMC Bioinformatics* 2013, **14 Suppl 11:**S1.
18. Sathirapongsasuti JF, Lee H, Horst BA, Brunner G, Cochran AJ, Binder S, Quackenbush J, Nelson SF: **Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV.** *Bioinformatics* 2011, **27:**2648-2654.
19. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK: **VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing.** *Genome Res* 2012, **22:**568-576.
20. Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, Janoueix-Lerosey I, Delattre O, Barillot E: **Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data.** *Bioinformatics* 2012, **28:**423-425.
21. Amarasinghe KC, Li J, Halgamuge SK: **CoNVEX: copy number variation estimation in exome sequencing data using HMM.** *BMC Bioinformatics* 2013, **14 Suppl 2:**S2.
22. Magi A, Tattini L, Cifola I, D'Aurizio R, Benelli M, Mangano E, Battaglia C, Bonora E, Kurg A, Seri M, Magini P, Giusti B, Romeo G, Pippucci T, De Bellis G, Abbate R, Gensini GF: **EXCAVATOR: detecting copy number variants from whole-exome sequencing data.** *Genome Biol* 2013, **14:**R120.
23. Tonon G, Wong KK, Maulik G, Brennan C, Feng B, Zhang Y, Khatry DB, Protopopov A, You MJ, Aguirre AJ, Martin ES, Yang Z, Ji H, Chin L,

Depinho RA: **High-resolution genomic profiles of human lung cancer.** *Proc Natl Acad Sci U S A* 2005, **102:**9625-9630.
24. Kabbarah O, Nogueira C, Feng B, Nazarian RM, Bosenberg M, Wu M, Scott KL, Kwong LN, Xiao Y, Cordon-Cardo C, Granter SR, Ramaswamy S, Golub T, Duncan LM, Wagner SN, Brennan C, Chin L: **Integrative genome comparison of primary and metastatic melanomas.** *PLoS One* 2010, **5:**e10770.
25. Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, Stebbings LA, Leroy C, Edkins S, Hardy C, Teague JW, Menzies A, Goodhead I, Turner DJ, Clee CM, Quail MA, Cox A, Brown C, Durbin R, Hurles ME, Edwards PA, Bignell GR, Stratton MR, Futreal PA: **Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing.** *Nat Genet* 2008, **40:**722-729.
26. Zhang C, Zhang C, Chen S, Yin X, Pan X, Lin G, Tan Y, Tan K, Xu Z, Hu P, Li X, Chen F, Xu X, Li Y, Zhang X, Jiang H, Wang W: **A single cell level based method for copy number variation analysis by low coverage massively parallel sequencing.** *PLoS One* 2013, **8:**e54236.
27. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics* 2010, **26:**841-842.
28. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloe D, Le Gall C, Schaeffer B, Le Crom S, Guedj M, Jaffrezic F: **A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis.** *Brief Bioinform* 2012, **17:**17.
29. Ioannidis JP, Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, Falchi M, Furlanello C, Game L, Jurman G, Mangion J, Mehta T, Nitzberg M, Page GP, Petretto E, van Noort V: **Repeatability of published microarray gene expression analyses.** *Nat Genet* 2009, **41:**149-155.
30. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.** *Bioinformatics* 2003, **19:**185-193.
31. **Gene Expresion Omnibus [** http://www.ncbi.nlm.nih.gov/geo **].**
32. Yoshida K, Sanada M, Shiraishi Y, Nowak D, Nagata Y, Yamamoto R, Sato Y, Sato-Otsubo A, Kon A, Nagasaki M, Chalkidis G, Suzuki Y, Shiosaka M, Kawahata R, Yamaguchi T, Otsu M, Obara N, Sakata-Yanagimoto M, Ishiyama K, Mori H, Nolte F, Hofmann WK, Miyawaki S, Sugano S, Haferlach C, Koeffler HP, Shih LY, Haferlach T, Chiba S, Nakauchi H, et al: **Frequent pathway mutations of splicing machinery in myelodysplasia.** *Nature* 2011, **478:**64-69.
33. **European Genome-phenome Archive [https://**http://www.ebi.ac.uk/ega/ **].**
34. Iannelli F, Collino A, Sinha S, Radaelli E, Nicoli P, D'Antiga L, Sonzogni A, Faivre J, Annick Buendia M, Sturm E, Thompson RJ, Knisely AS, Natoli G, Ghisletti S, Ciccarelli FD: **Massive gene amplification drives paediatric hepatocellular carcinoma caused by bile salt export pump deficiency.** *Nat Commun* 2014, **5:**3850.
35. Van Loo P, Nordgard SH, Lingjaerde OC, Russnes HG, Rye IH, Sun W, Weigman VJ, Marynen P, Zetterberg A, Naume B, Perou CM, Borresen-Dale AL, Kristensen VN: **Allele-specific copy number analysis of tumors.** *Proc Natl Acad Sci U S A* 2010, **107:**16910-16915.

36. **Sequence Read Archive [** http://www.ncbi.nlm.nih.gov/sra **].**
37. **Novoalign [** http://novocraft.com **].**
38. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25:**1754-1760.
39. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.
40. Mosen-Ansorena D, Aransay AM, Rodriguez-Ezpeleta N: **Comparison of methods to detect copy number alterations in cancer using simulated and real genotyping data.** *BMC Bioinformatics* 2012, **13:**192.
41. Parisi F, Ariyan S, Narayan D, Bacchiocchi A, Hoyt K, Cheng E, Xu F, Li P, Halaban R, Kluger Y: **Detecting copy number status and uncovering subclonal markers in heterogeneous tumor biopsies.** *BMC Genomics* 2011, **12:**230.
42. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR: **A census of human cancer genes.** *Nat Rev Cancer* 2004, **4:**177-183.
43. An O, Pendino V, D'Antonio M, Ratti E, Gentilini M, Ciccarelli FD: **NCG 4.0: the network of cancer genes in the era of massive mutational screenings of cancer genomes.** *Database (Oxford)* 2014, **2014:**bau015.

Figures

Figure 1 - GeneCNV workflow.

GeneCNV uses whole exome sequencing to calculate the coverage of all targeted genes in test (*e.g* tumor) and reference (*e.g* matched normal) exomes independently (A). It normalizes the gene coverage within and between exomes using median normalization and quantile normalization, respectively (B). It then measures the $\log_2$ratio of gene coverage ($L2R_{GC}$) between tumor and matched normal exomes for each targeted gene (C) and identifies the regions of allelic balance in the tumor (D). Within allelic balanced regions, GeneCNV further identifies regions that maintain a diploid status in the tumor (E) and defines $L2R_{GC}$ thresholds for amplifications and deletions from them (F). This allows the detection of amplified (green), deleted (red), CN-LOH (yellow) and two copy (grey) genes along the whole exome (G). The parameters to define diploid regions and the sample-specific thresholds are empirical and can be all modified by the user.

Figure 2 - Performance comparison of GeneCNV with other exome-based methods.

Reported are sensitivity, specificity, accuracy and Jaccard index of the four exome-based methods (GeneCNV, ExomeCNV, VarScan 2, and EXCAVATOR) in detecting all CNVs in the 28 samples as compared to SNP array (A); the trade-off between sensitivity and specificity of each method (B); performance of the four methods in detecting amplified genes (C), deleted genes (D), CN-LOH genes and clonal variant genes (E). For CN-LOHs assessment, only GeneCNV and ExomeCNV are compared because VarScan 2, and EXCAVATOR cannot detect CN-LOHs.

Figure 3 - GeneCNV graphical output.

Shown are exemplar graphical outputs of GeneCNV on a single tumor (sample MDS-09, A-D) and on a cohort of tumors (all 22 myleodysplasia exomes, E-G, see also Additional File 1). In the case of the single sample analysis, provided are the plots of gene coverage in the tumor and the normal exomes before and after normalizations (A); the tumor $L2R_{GC}$ spectrum with amplified, deleted and CN-LOH genes highlighted in colors (B); the percentage of somatically amplified, deleted and CN-LOH genes in each tumor chromosome and in regions representing 10% of chromosome arms (C); and the circos plot summarizing all alterations and reporting known [42] and candidate [43] altered cancer genes (D). In the case of multiple sample comparison, shown are the percentage of amplified, deleted and CN-LOH genes in each tumor exome (E); the number of genes whose modifications recur across samples (F); and the circos plot with all genomic alterations and recurrently altered cancer genes [42] (G).

Additional files

Additional file 1 – Samples used for the comparison of exome-based methods

This file contains Supplementary Table S1 with the description of the 28 samples used for method comparison. The file is provided in XLS format.

Additional file 2 – Amplified, deleted, and CN-LOH genes in the 28 tumor exomes.

This file contains Supplementary Table S2 and reports the copy number status of all targeted genes in each of the 28 tumor exomes, as detected by ASCAT in the SNP array and by the four exome-based methods. The file is provided in XLS format.

Additional file 3 – Performance of the four exome-based methods in detecting variant genes

This file contains Supplementary Table S3 and reports sensitivity, specificity, accuracy, and Jaccard index of all four methods in identifying variant genes in 28 tumor exomes. The file is provided in XLS format.

Additional file 4 – Performance of the four exome-based methods in detecting amplified and deleted genes

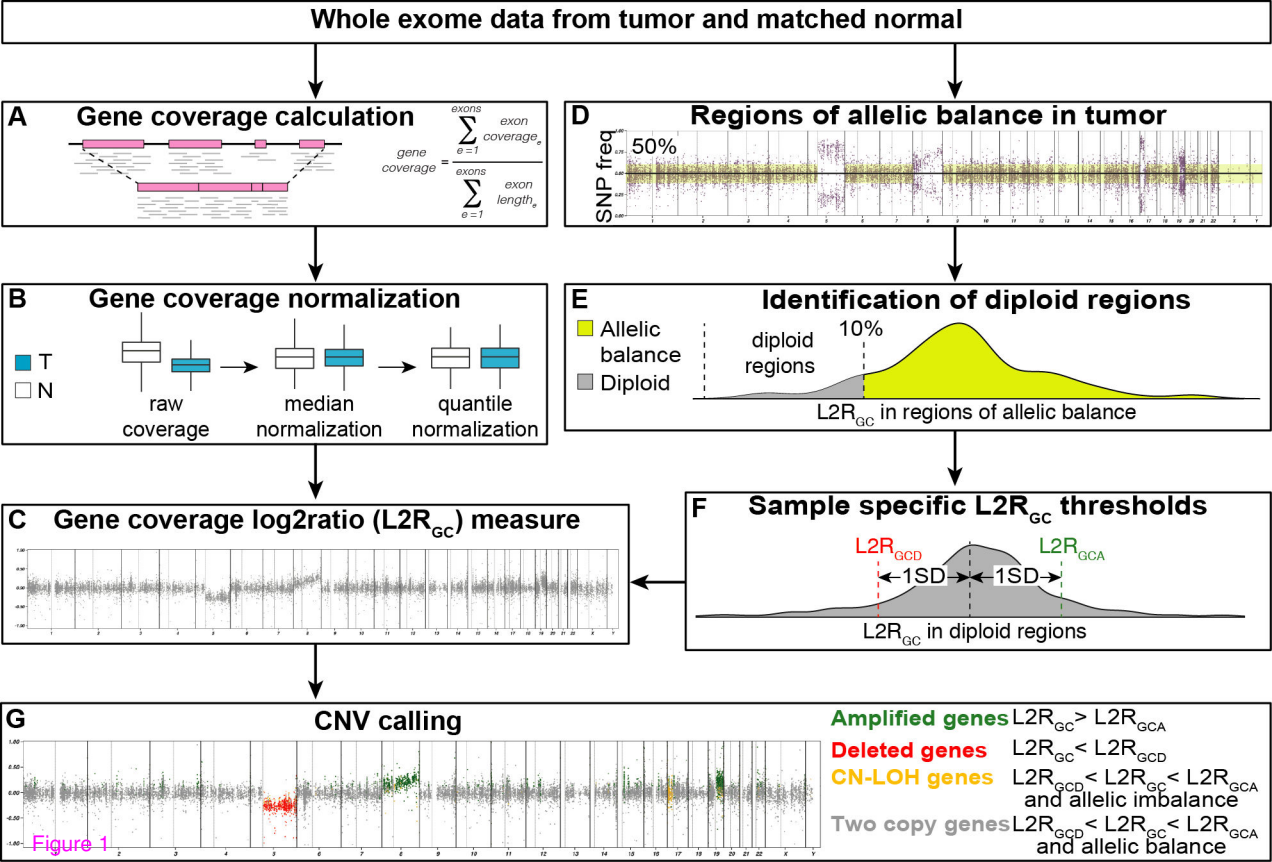
This file contains Supplementary Figure S1 and reports sensitivity, specificity, accuracy, and Jaccard index of all four methods in detecting amplified and deleted genes in the 28 tumor exomes. The file is provided in PDF format.

Additional file 5 – Clonally amplified, deleted, and CN-LOH genes in the 28 tumor exomes

This file contains Supplementary Table S4 and reports only clonal events occurring in each of the 28 tumor exomes. The file is provided in XLS format.

Additional file 6 – Performance of the four exome-based methods in detecting clonal events

This file contains Supplementary Table S5 and reports sensitivity, specificity, accuracy, and Jaccard index of all four methods in identifying clonal modifications occurring in each of the 28 tumor exomes. The file is provided in XLS format.


Whole exome data from tumor and matched normal
A
Gene coverage calculation
gene coverage = Σ exon coverage_e (e=1 to exons) / Σ exon length_e (e=1 to exons)
D
Regions of allelic balance in tumor
SNP freq
50%
B
Gene coverage normalization
T
N
raw coverage
median normalization
quantile normalization
E
Identification of diploid regions
Allelic balance
Diploid
diploid regions
10%
L2R_GC in regions of allelic balance
C
Gene coverage log2ratio (L2R_GC) measure
F
Sample specific L2R_GC thresholds
L2R_GCD
L2R_GCA
1SD
1SD
L2R_GC in diploid regions
G
CNV calling
Amplified genes L2R_GC > L2R_GCA
Deleted genes L2R_GC < L2R_GCD
CN-LOH genes L2R_GCD < L2R_GC < L2R_GCA and allelic imbalance
Two copy genes L2R_GCD < L2R_GC < L2R_GCA and allelic balance
Figure 1

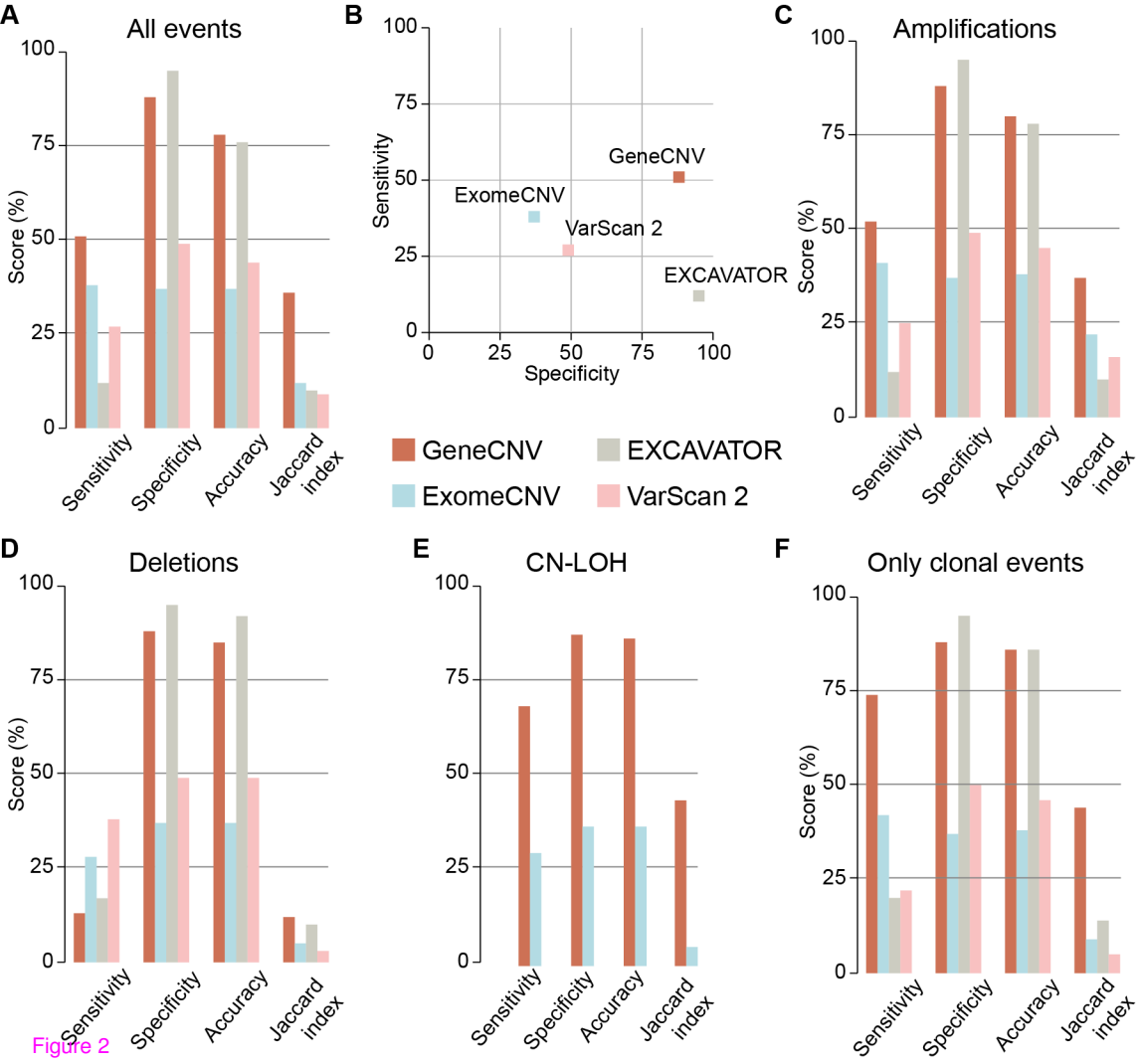

A
All events
Score (%)
100
75
50
25
0
Sensitivity
Specificity
Accuracy
Jaccard index
B
Sensitivity
Specificity
100
75
50
25
0
0
25
50
75
100
GeneCNV
ExomeCNV
VarScan 2
EXCAVATOR
C
Amplifications
Score (%)
Sensitivity
Specificity
Accuracy
Jaccard index
GeneCNV
EXCAVATOR
ExomeCNV
VarScan 2
D
Deletions
Score (%)
Sensitivity
Specificity
Accuracy
Jaccard index
E
CN-LOH
Sensitivity
Specificity
Accuracy
Jaccard index
F
Only clonal events
Score (%)
Sensitivity
Specificity
Accuracy
Jaccard index

Figure 2

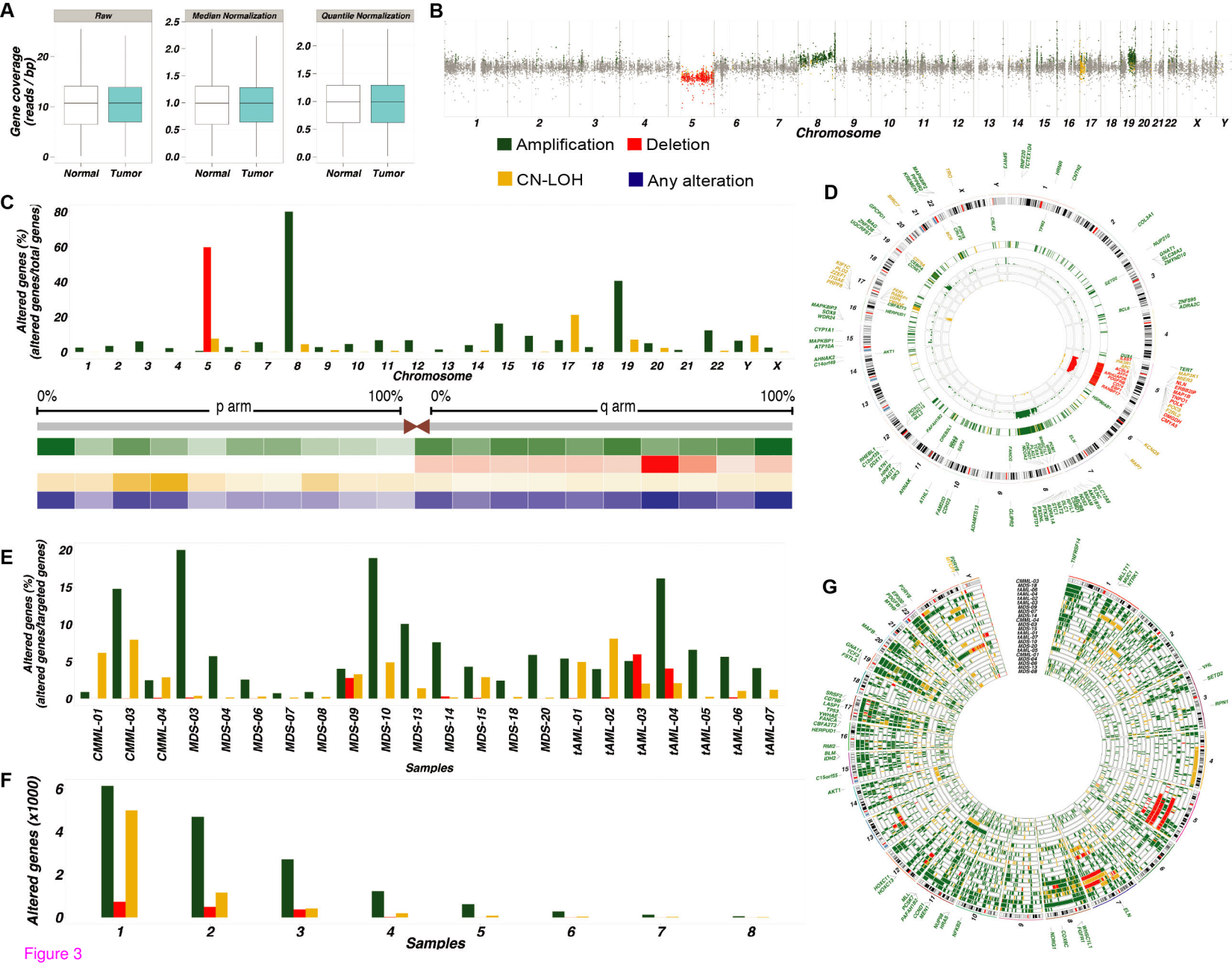

A
Raw
Median Normalization
Quantile Normalization
Gene coverage (reads / bp)
Normal
Tumor
B
Chromosome
Amplification
Deletion
CN-LOH
Any alteration
C
Altered genes (%) (altered genes/total genes)
Chromosome
0%
p arm
100%
0%
q arm
100%
D
E
Altered genes (%) (altered genes/targeted genes)
CMML-01
CMML-03
CMML-04
MDS-03
MDS-04
MDS-06
MDS-07
MDS-08
MDS-09
MDS-10
MDS-13
MDS-14
MDS-15
MDS-18
MDS-20
tAML-01
tAML-02
tAML-03
tAML-04
tAML-05
tAML-06
tAML-07
Samples
F
Altered genes (x1000)
Samples
G

Figure 3

Additional files provided with this submission:

Additional file 1: Sinha_AdditionalFile1.xlsx, 47K
http://genomebiology.com/imedia/6199596331420146/supp1.xlsx
Additional file 2: Sinha_AdditionalFile2.xlsx, 49K
http://genomebiology.com/imedia/1825669148142014/supp2.xlsx
Additional file 3: Sinha_AdditionalFile3.xlsx, 43K
http://genomebiology.com/imedia/1721990715142014/supp3.xlsx
Additional file 4: Sinha_AdditionalFile4.pdf, 676K
http://genomebiology.com/imedia/1455992140142014/supp4.pdf
Additional file 5: Sinha_AdditionalFile5.xlsx, 47K
http://genomebiology.com/imedia/1061629542142014/supp5.xlsx
Additional file 6: Sinha_AdditionalFile6.xlsx, 44K
http://genomebiology.com/imedia/4766091131420146/supp6.xlsx